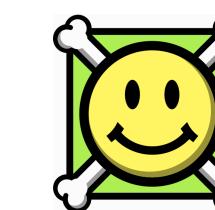




Lei Wang (王磊)
Institute of Physics, CAS



Zhendong Cao, Xiaoshan Luo,
Jian Lv, and LW, 2403.15734



[deepmodeling/CrystalFormer](https://github.com/deepmodeling/CrystalFormer)

SQAI-NCTS Workshop on Tensor Network and Quantum Embedding

March 25-29, 2024

Hongo campus, The University of Tokyo, Tokyo, Japan

About

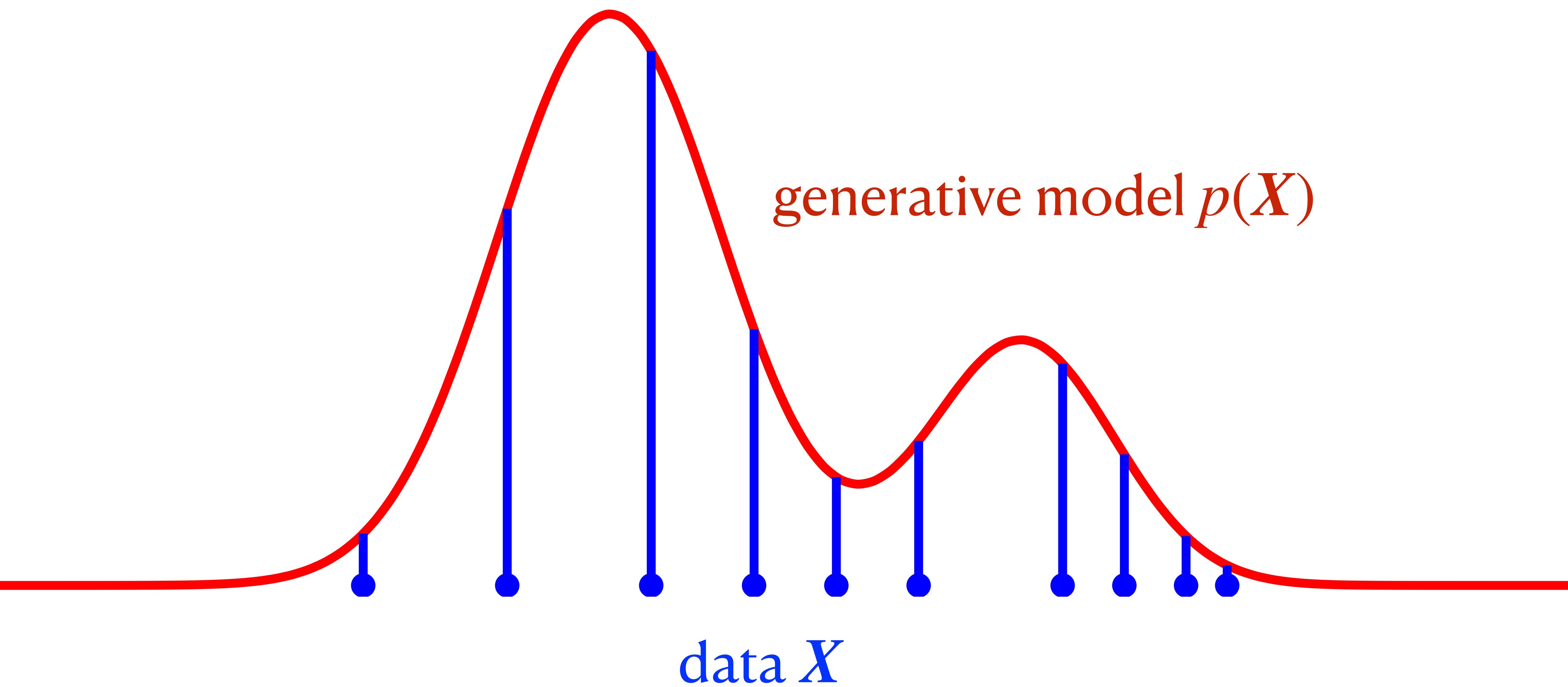
This workshop focuses on the related topics of tensor network and its potential application in the wide research field such as the quantum algorithm. Topics covered by it include the quantum embedding, quantum machine learning, condensed matter problems, etc.

Dates

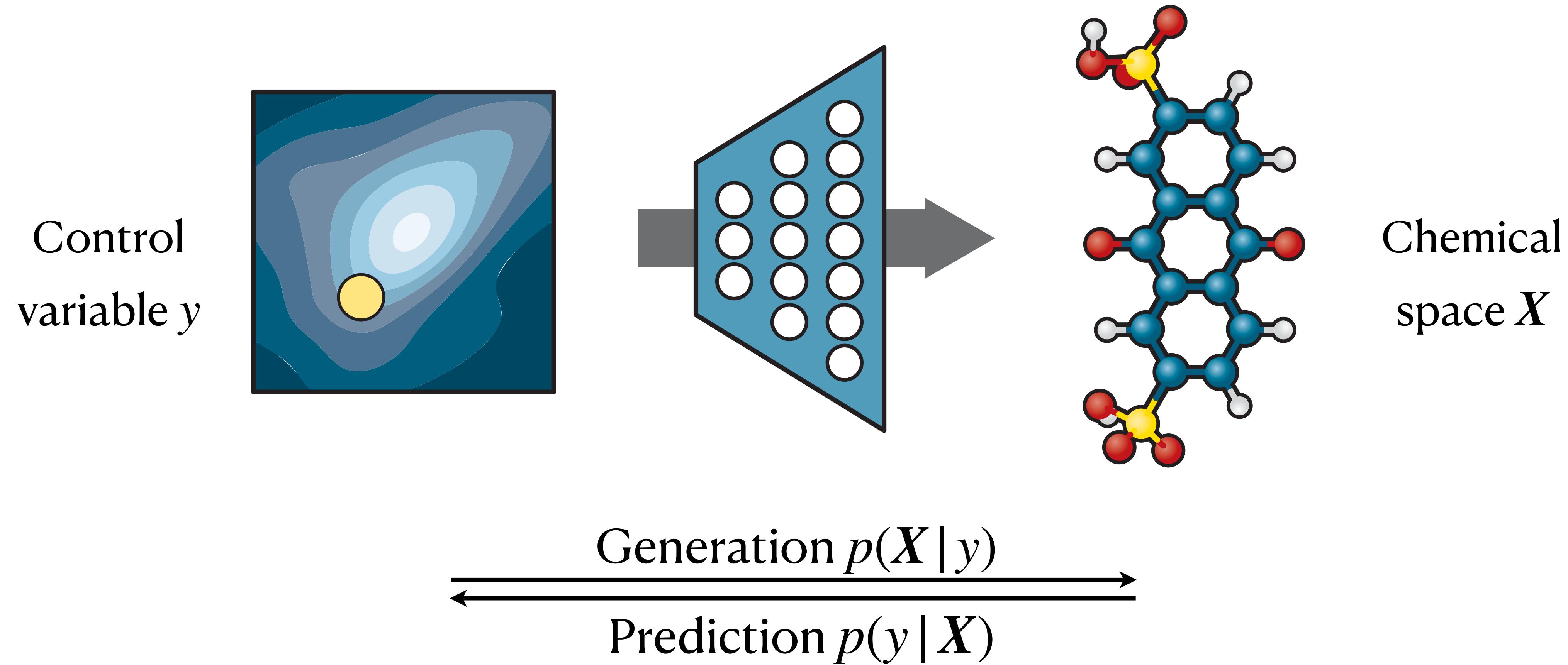
March 25 - 29, 2024

Venue

Hongo campus, the University of Tokyo, Tokyo, Japan



Generative AI for matter engineering



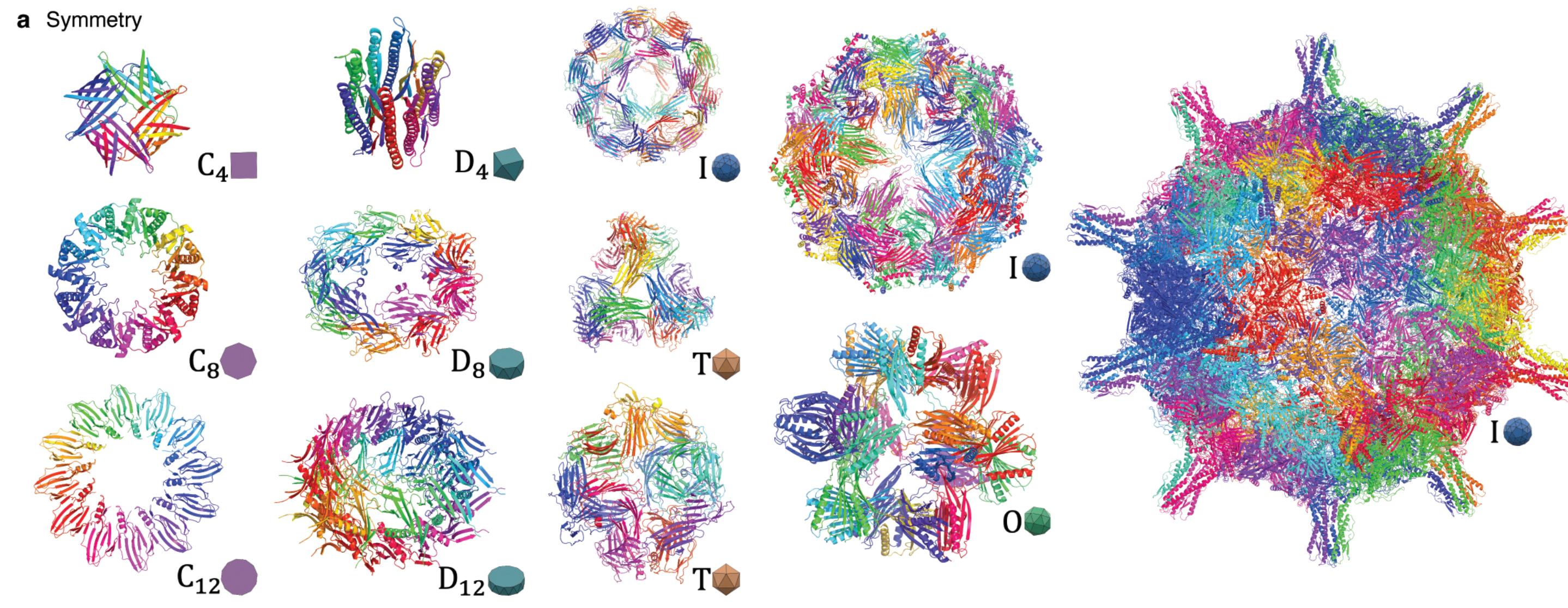
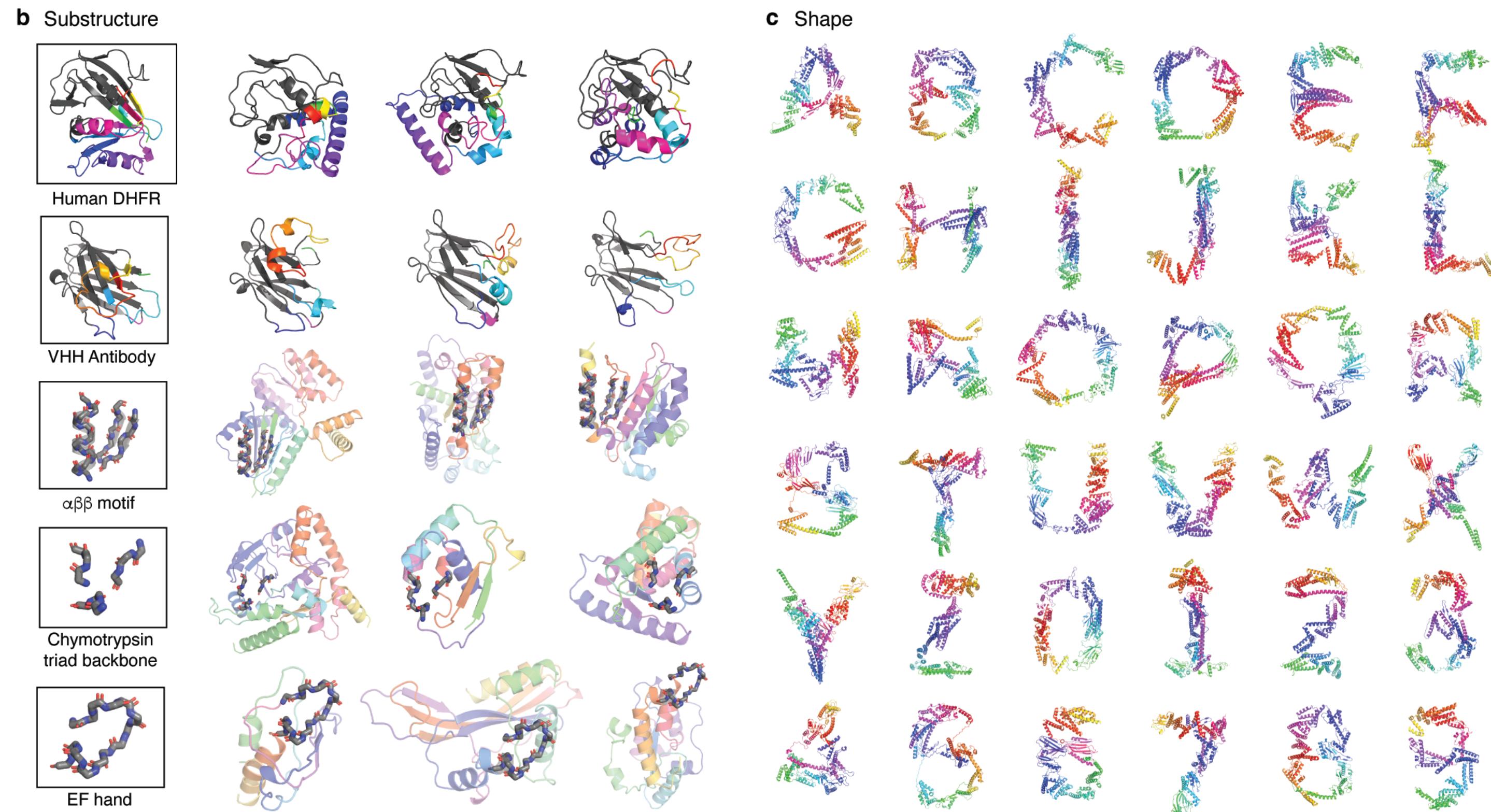
Input:

An astronaut riding a horse in photorealistic style.

Output:

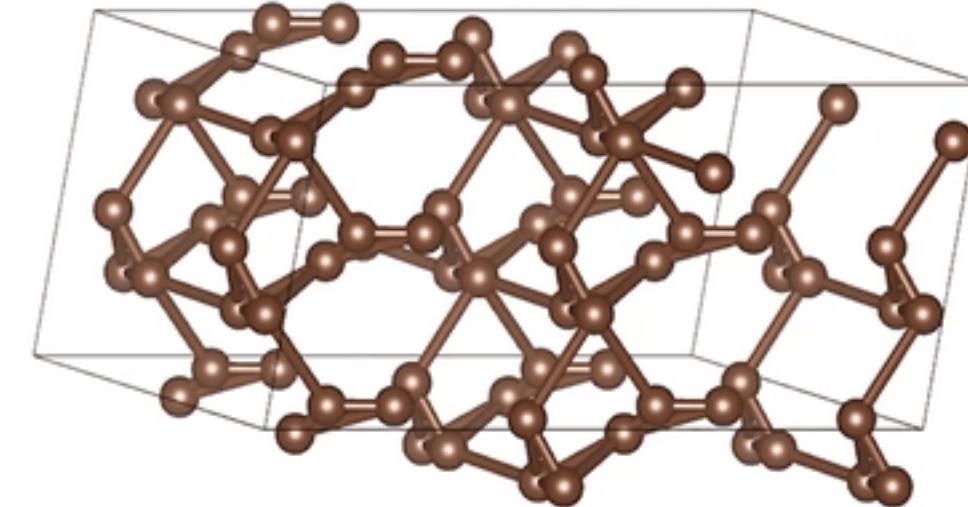


$\text{image} \sim p(\text{image} \mid \text{text})$
<https://openai.com/product/dall-e-3>

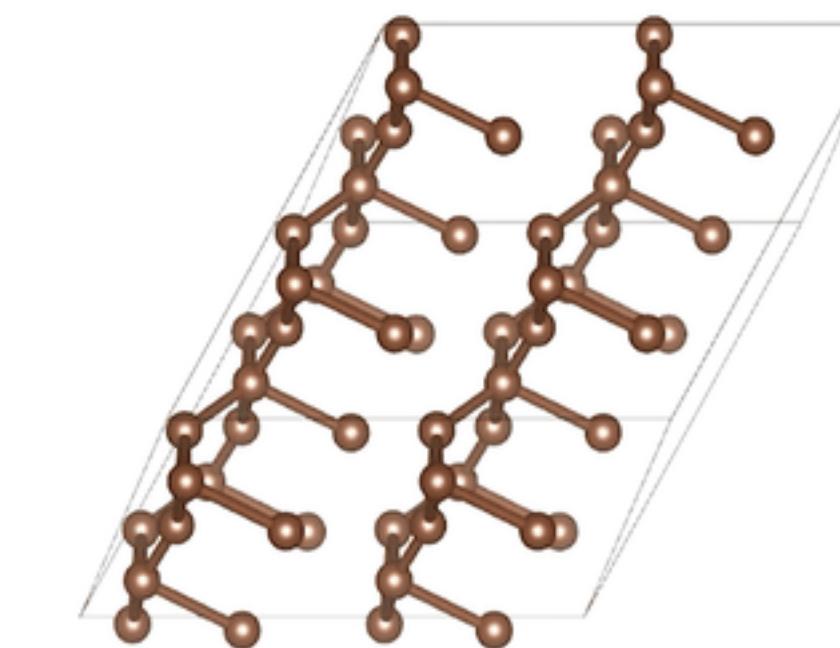

 $p(\text{protein} \mid \text{symmetry})$

 $p(\text{protein} \mid \text{substructure})$
 $p(\text{protein} \mid \text{shape})$

$p(\text{carbon allotropes} \mid \text{bandgap})$

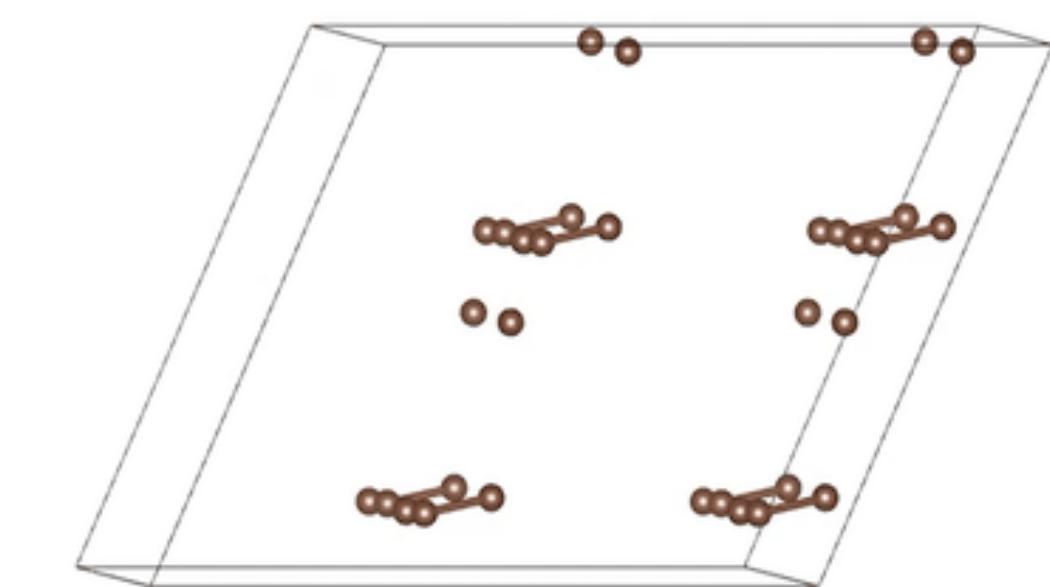
Band gap = 0 eV



2 eV



4 eV

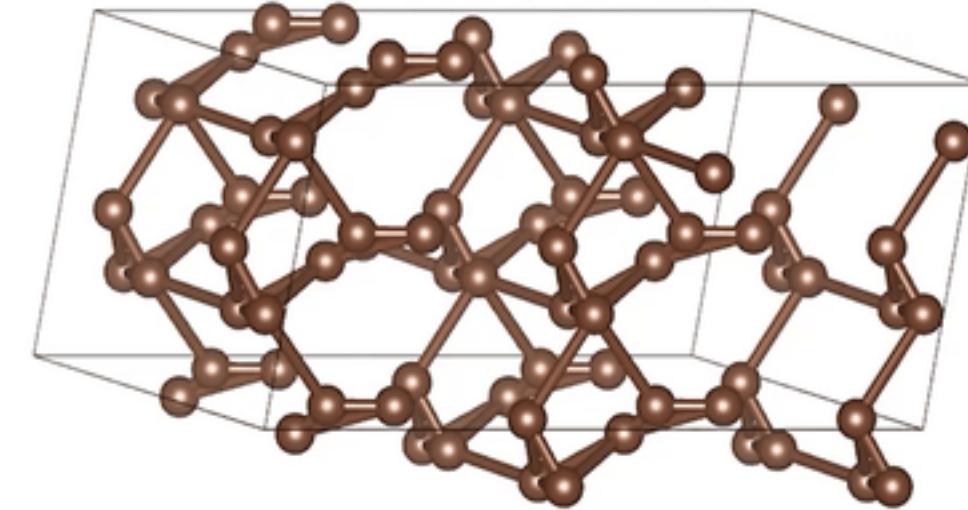


Zheng et al, 2306.05445, <https://distributionalgraphomer.github.io/>

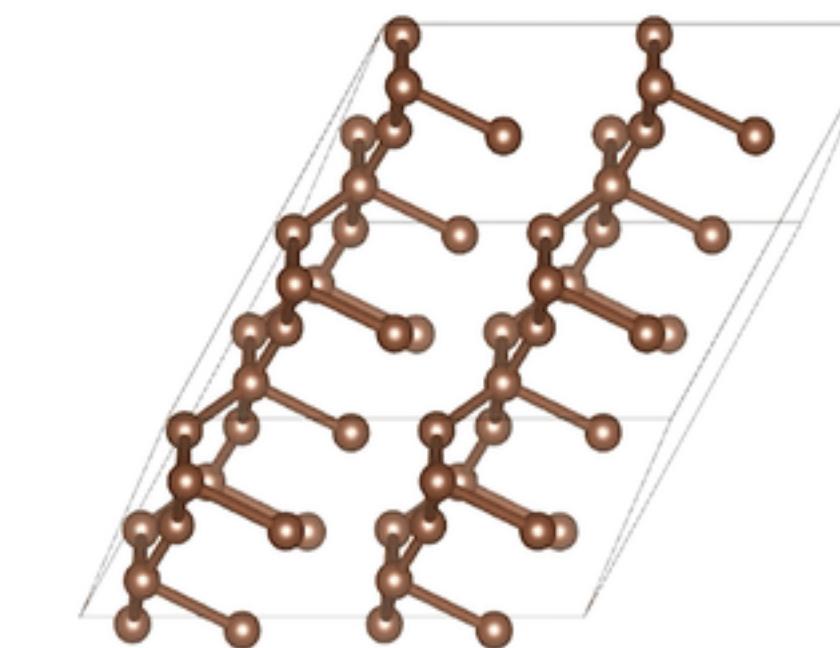
See also Xie et al, 2110.06197, Jiao et al, 2309.04475, Yang et al, 2311.09235, Zeni et al, 2312.03687, Jiao et al, 2402.03992 ...

$p(\text{carbon allotropes} \mid \text{bandgap})$

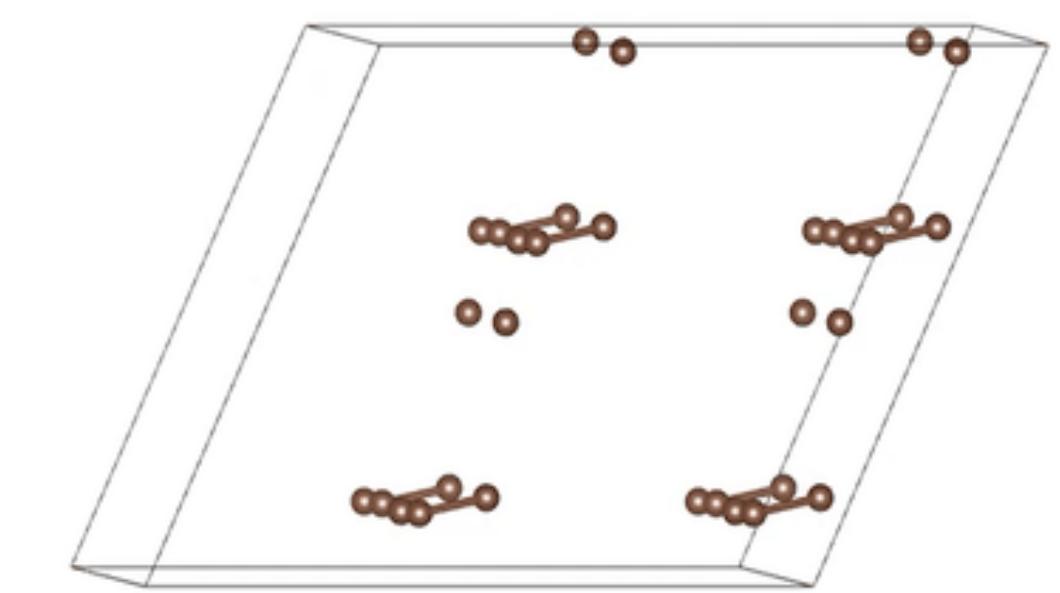
Band gap = 0 eV



2 eV



4 eV



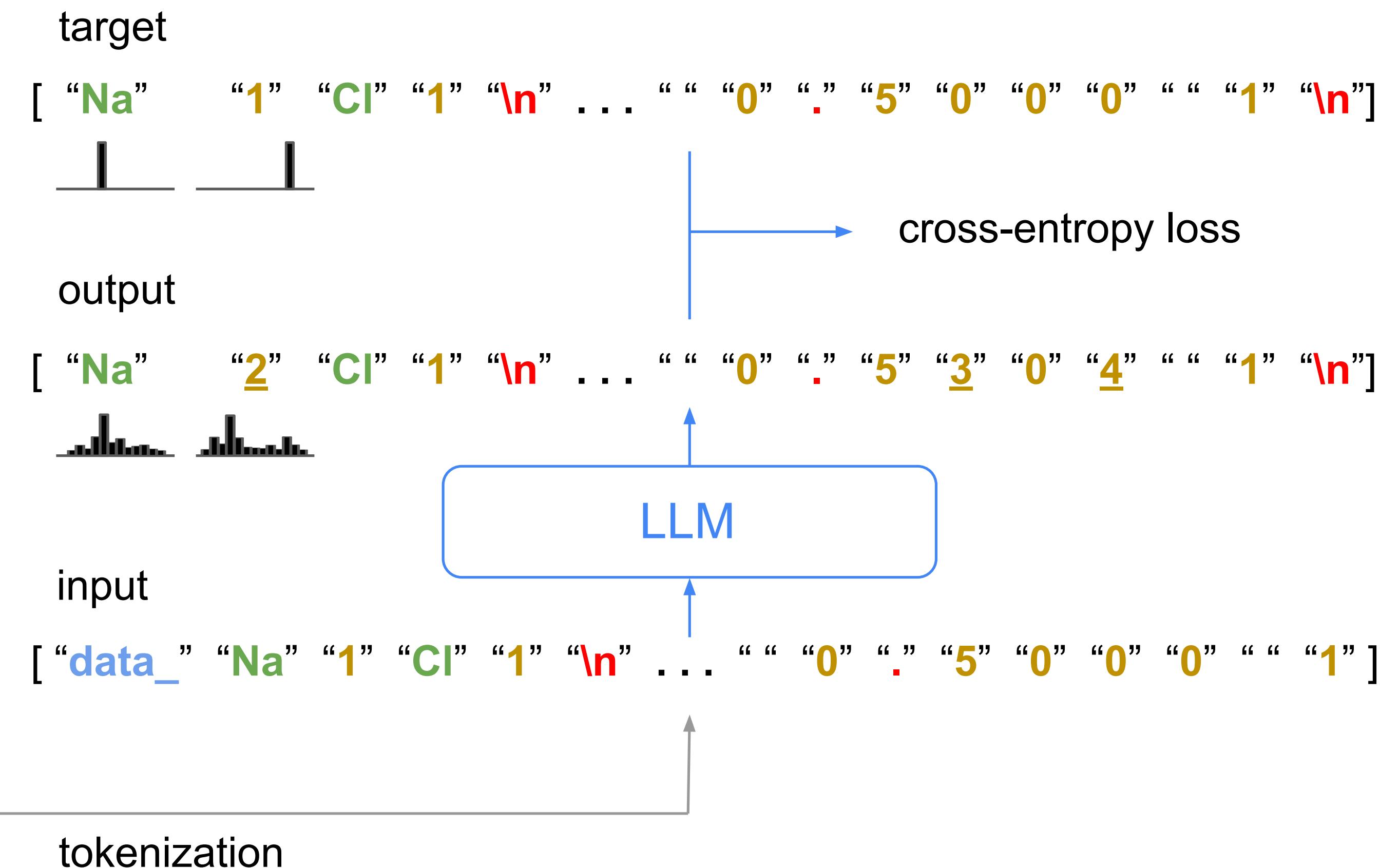
Zheng et al, 2306.05445, <https://distributionalgraphomer.github.io/>

See also Xie et al, 2110.06197, Jiao et al, 2309.04475, Yang et al, 2311.09235, Zeni et al, 2312.03687, Jiao et al, 2402.03992 ...

The large language model way

Crystallographic Information File

```
data_Na1Cl1
_symmetry_space_group_name_H-M 'P1'
_cell_length_a 3.9893
_cell_length_b 3.9893
_cell_length_c 3.9893
_cell_angle_alpha 60.0000
_cell_angle_beta 60.0000
_cell_angle_gamma 60.0000
_symmetry_Int_Tables_number 1
_chemical_formula_structural NaCl
_chemical_formula_sum 'Na1 Cl1'
_cell_volume 44.8931
_cell_formula_units_Z 1
loop_
_symmetry_equiv_pos_site_id
_symmetry_equiv_pos_as_xyz
 1 'x, y, z'
loop_
_atom_site_type_symbol
_atom_site_label
_atom_site_symmetry_multiplicity
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_occupancy
Cl Cl0 1 0.0000 0.0000 0.0000 1
Na Na1 1 0.5000 0.5000 0.5000 1
```



CrystaLLM

https://crystallm.com

Generate a crystal structure from a composition *

Composition: PbTe 3 optional Space group optional Generate!

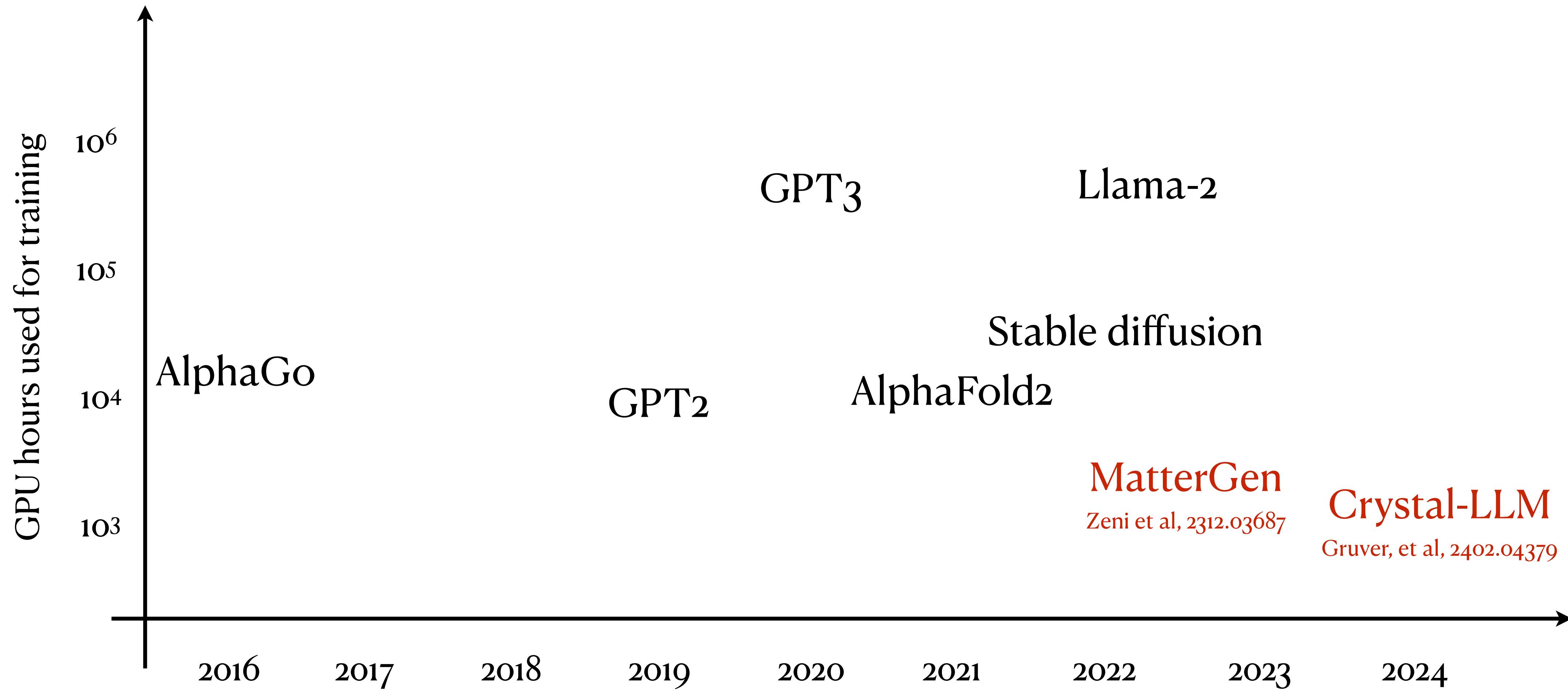
► Advanced options

Te Pb

R-3m $a=4.020\text{\AA}$ $b=4.020\text{\AA}$ $c=13.419\text{\AA}$ $\alpha=90.000^\circ$ $\beta=90.000^\circ$ $\gamma=120.000^\circ$

CrystaLLM v1.0

* This service is intended for non-commercial use only. We make no guarantees regarding the accuracy of the predictions. The predictions are made by machine learning models that are statistical in nature, and not necessarily based on physical theory. If you would like to use this



Is scale all you need ?

We have much less data



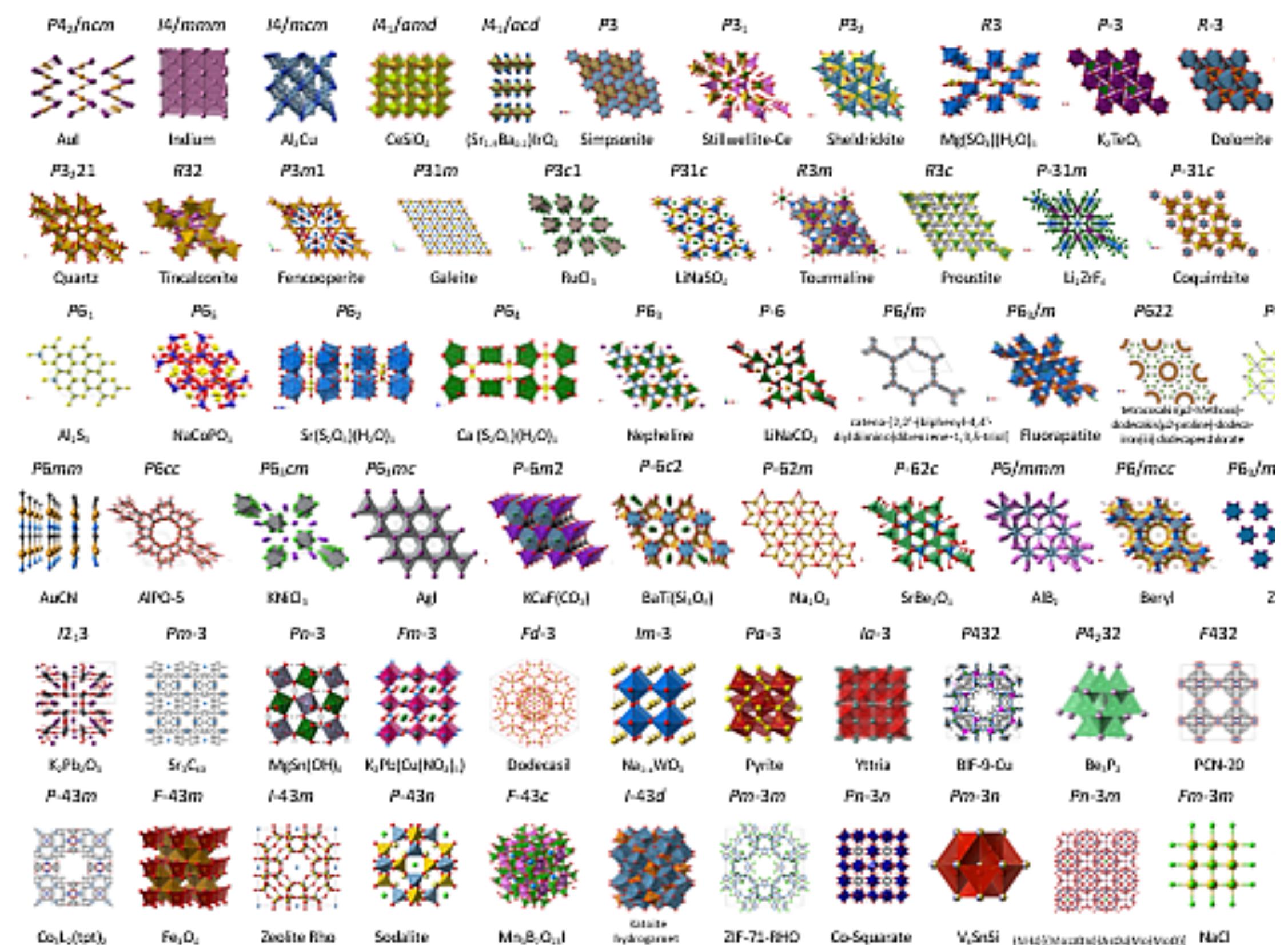
Over 250 billion pages



> 291,000 crystal structures

Data, compute, and model size need to scale simultaneously Kaplan et al, 2001.08361

Crystals \neq a bunch of words or pixels



more data and compute

more physics and symmetries

<http://www.incompleteideas.net/InclIdeas/BitterLesson.html>

The screenshot shows a web browser window with two tabs. The active tab is titled "The Bitter Lesson" and displays the text of the blog post. The text discusses the importance of computation in AI research, mentioning Moore's law and the exponential falling cost of computation. It contrasts this with the human-knowledge approach, noting that while it might work in the short term, it often leads to less effective long-term solutions.

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

<https://staff.fnwi.uva.nl/m.welling/wp-content/uploads/Model-versus-Data-AI-1.pdf>

The screenshot shows a web browser window with two tabs. The active tab is titled "The Bitter Lesson" and displays a quote from Max Welling. The quote discusses the polarizing views within the AI community regarding the need for models versus data and computation. It highlights the work of Rich Sutton and others who advocate for compute-driven AI.

Do we still need models or just more data and compute?

Max Welling, Amsterdam, April 20 2019

This question, or versions of it, seems to divide the AI community. And much like Bayesians and Frequentists they hold rather strong polarizing views on the matter. The question seems to come in different flavors: symbolic AI or statistical AI, white box AI or black box AI, model driven or data driven AI, generative or discriminative AI? A recent blog by Rich Sutton adds to the list compute-driven AI versus human-knowledge based AI. The discussion is both fascinating and deeply fundamental. We should all be thinking about these questions.

Let me start by stating that I am a strong believer in the power of computation and its relevance to ML. One of the reasons I have a part-time position at Qualcomm is precisely because I believe one of the fastest ways to make progress in AI is to make specialized hardware for AI computations. I am also a strong proponent of deep learning. Much of my research portfolio is currently based on DL. I think it is the biggest hammer we have produced thus far and I witness its impact every day. (So please, Gary Marcus, do not write after reading this post that I am an opponent of DL).

In the blog by Rich Sutton, "The Bitter Lesson", one can read something of the kind: one should work on scalable methods that can maximally leverage compute and forget about modeling the world. A number of examples are explained to support this claim, namely Deep Blue and AlphaGO who leverage search and learning rather than human strategies, speech recognition, visual object recognition etc. And we can add a few more to the list, melanoma detection and tumor detection, statistical machine translation etc. There is no doubt a trend here that cannot be ignored.

more data and compute

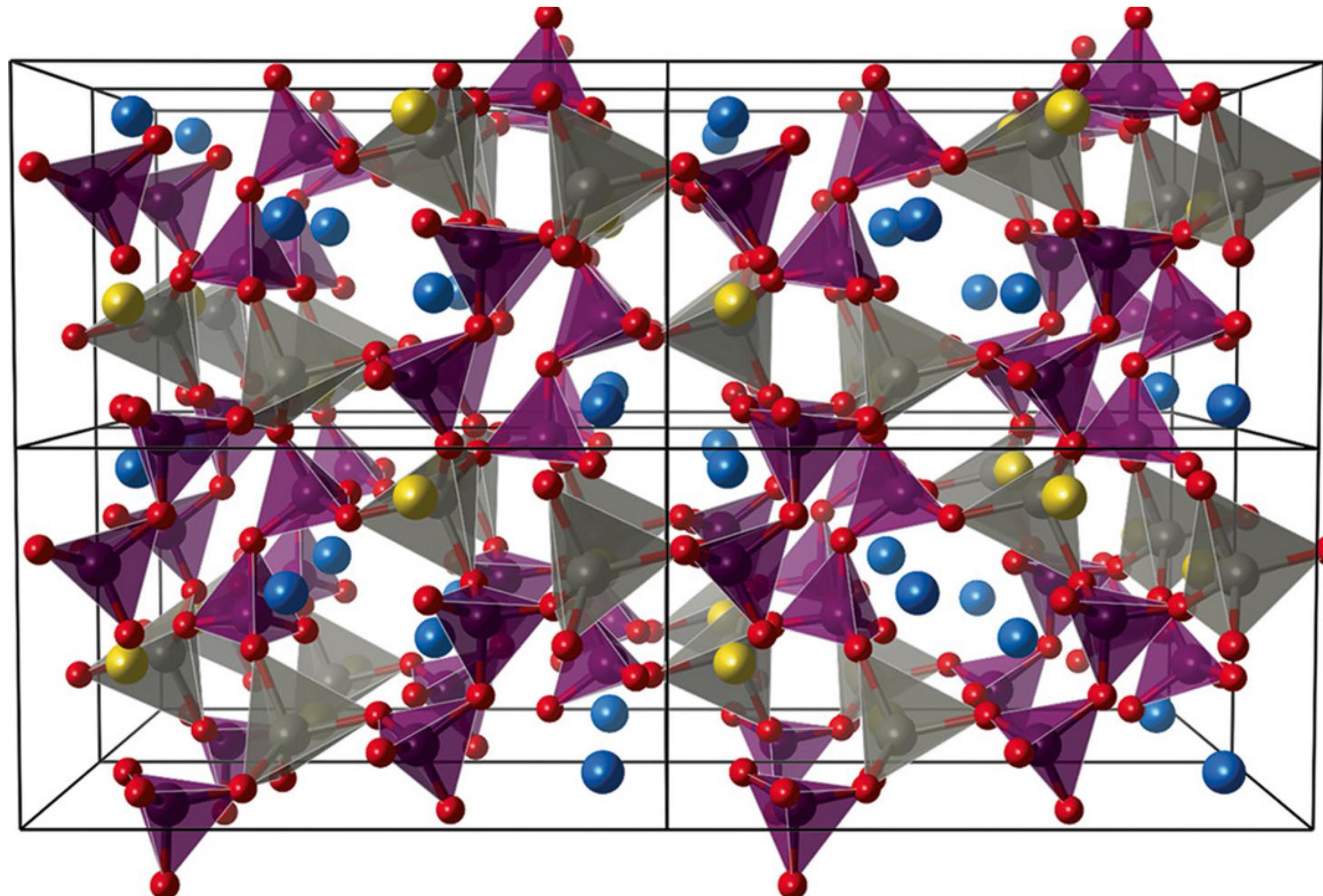
more physics and symmetries



How are crystals different from
languages/images/proteins ?

Space group symmetry

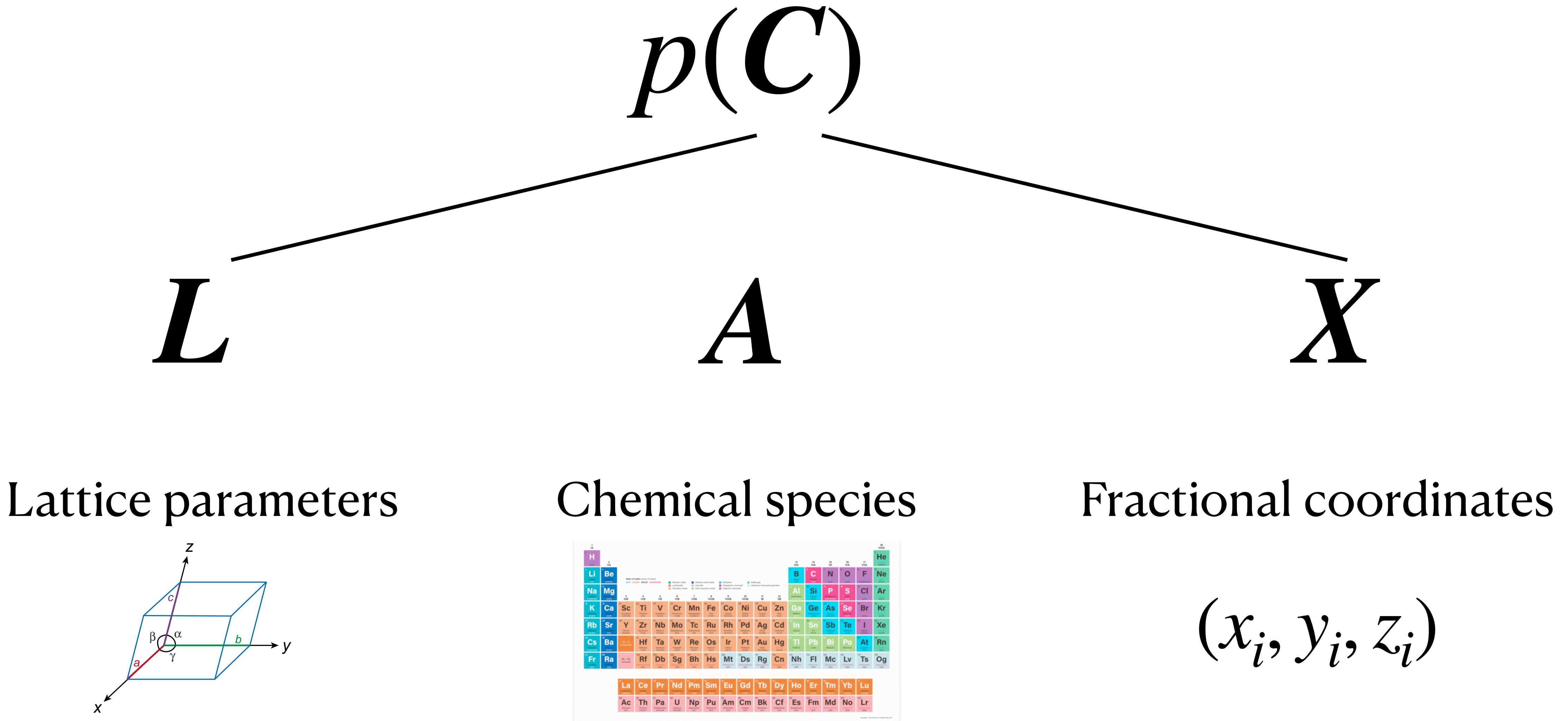
Classification of isometric transformations ($W^*x + b$) that keep the crystal unchanged



Translation
Mirror
Rotation
Inversion
Glide
Screw

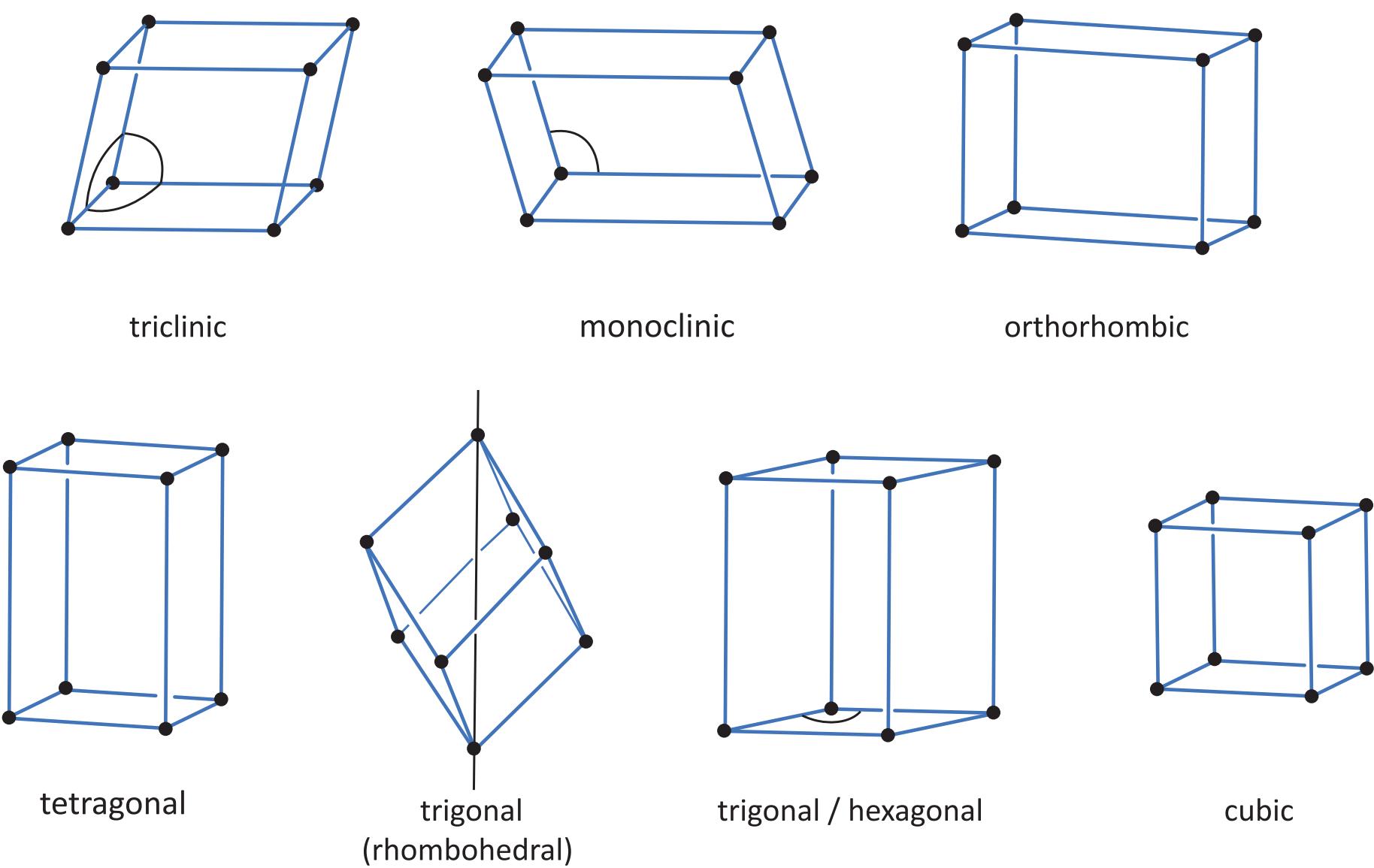
Only 230 space groups in three dimension!

Crystal generative model



Restrictions to L : crystal systems

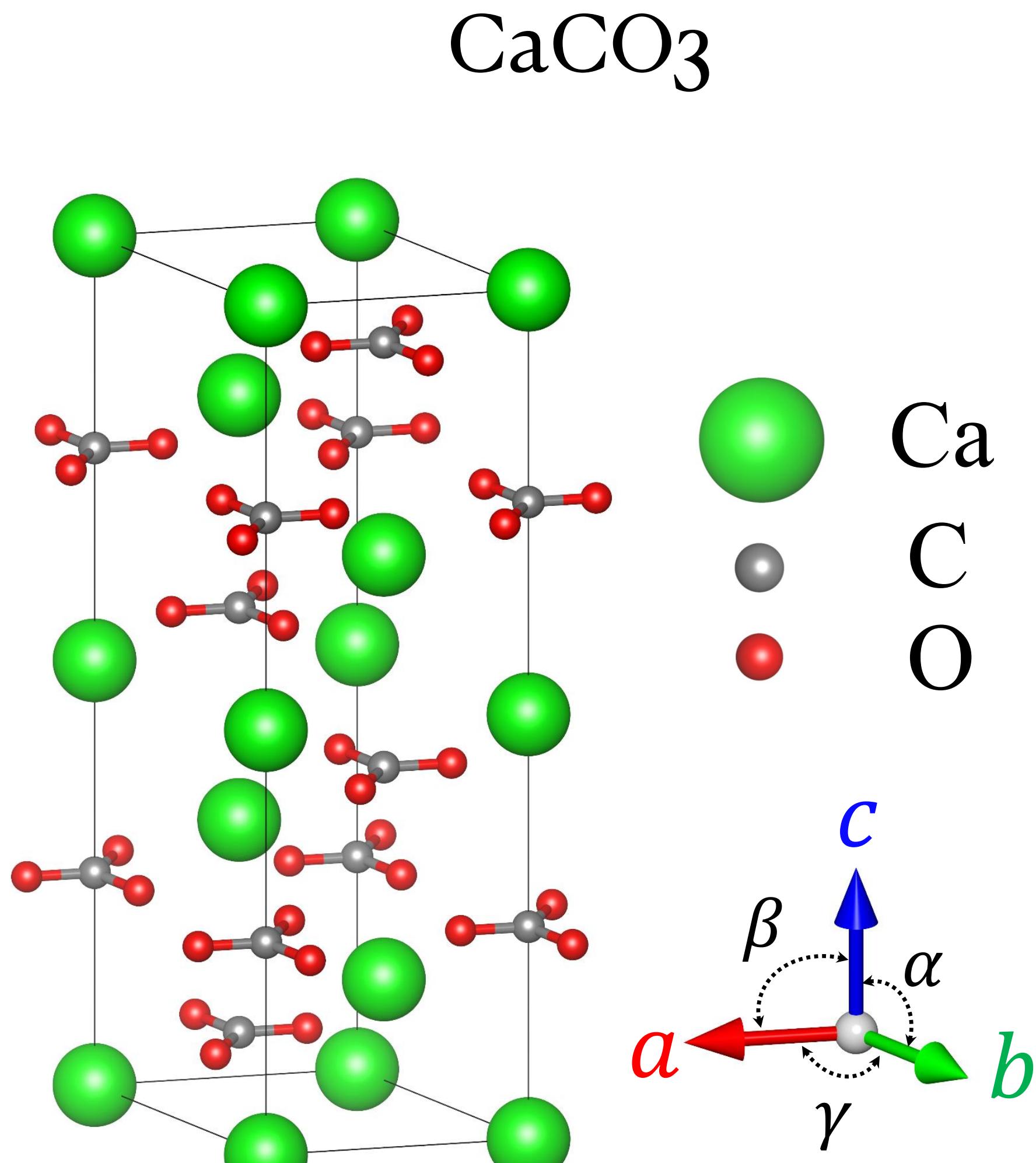
Crystal system		Restrictions concerning the	
		Length of axes	Angles of the cell
Triclinic	1-2	None ^a	None ^a
Monoclinic	3-15	None ^a	$\alpha = \gamma = 90^\circ$
Orthorhombic	16-74	None ^a	$\alpha = \beta = \gamma = 90^\circ$
Tetragonal	75-142	$a = b$	$\alpha = \beta = \gamma = 90^\circ$
Trigonal	143-167	$a = b$	$\alpha = \beta = 90^\circ, \gamma = 120^\circ$
Hexagonal	168-194	$a = b$	$\alpha = \beta = 90^\circ, \gamma = 120^\circ$
Cubic	195–230	$a = b = c$	$\alpha = \beta = \gamma = 90^\circ$



Restrictions to A and X : Wyckoff positions

Wyckoff Positions of Group R-3c (No. 167) [hexagonal axes]

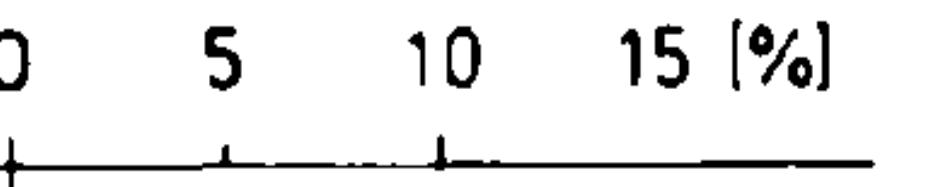
Multiplicity	Wyckoff letter	Fractional coordinates
		$(0, 0, 0) + (\frac{2}{3}, \frac{1}{3}, \frac{1}{3}) + (\frac{1}{3}, \frac{2}{3}, \frac{2}{3}) +$
36	f	$(x, y, z), (-y, x - y, z),$ $(-x + y, -x, z), (y, x, -z + \frac{1}{2}),$ $(x - y, -y, -z + \frac{1}{2}), (-x, -x + y, -z + \frac{1}{2}),$ $(-x, -y, -z), (y, -x + y, -z),$ $(x - y, x, -z), (-y, -x, z + \frac{1}{2}),$ $(-x + y, y, z + \frac{1}{2}), (x, x - y, z + \frac{1}{2})$
18	e	$(x, 0, \frac{1}{4}), (0, x, \frac{1}{4}),$ $(-x, -x, \frac{1}{4}), (-x, 0, \frac{3}{4}),$ $(0, -x, \frac{3}{4}), (x, x, \frac{3}{4})$
18	d	$(\frac{1}{2}, 0, 0), (0, \frac{1}{2}, 0),$ $(\frac{1}{2}, \frac{1}{2}, 0), (0, \frac{1}{2}, \frac{1}{2}),$ $(\frac{1}{2}, 0, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$
12	c	$(0, 0, z), (0, 0, -z + \frac{1}{2}),$ $(0, 0, -z), (0, 0, z + \frac{1}{2})$
6	b	$(0, 0, 0), (0, 0, \frac{1}{2})$
6	a	$(0, 0, \frac{1}{4}), (0, 0, \frac{3}{4})$



Wyckoff Positions of Group *P1* (No. 1)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
1	a	1	(x,y,z)

P1 is rare! Nature loves symmetric structures



Pnma

Fm-3m

C2/c

C2/m

Fd-3m

R-3m

Cmcm

Pm-3m

R3

P6/mmm

P21/m

R3c

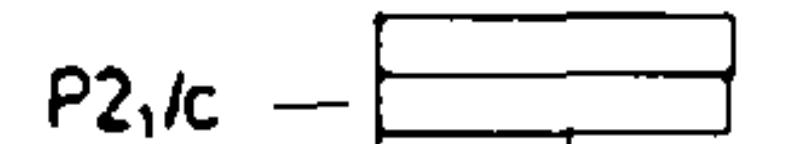
P63/m

P212121

Pna21

F43m

I4/mcm-



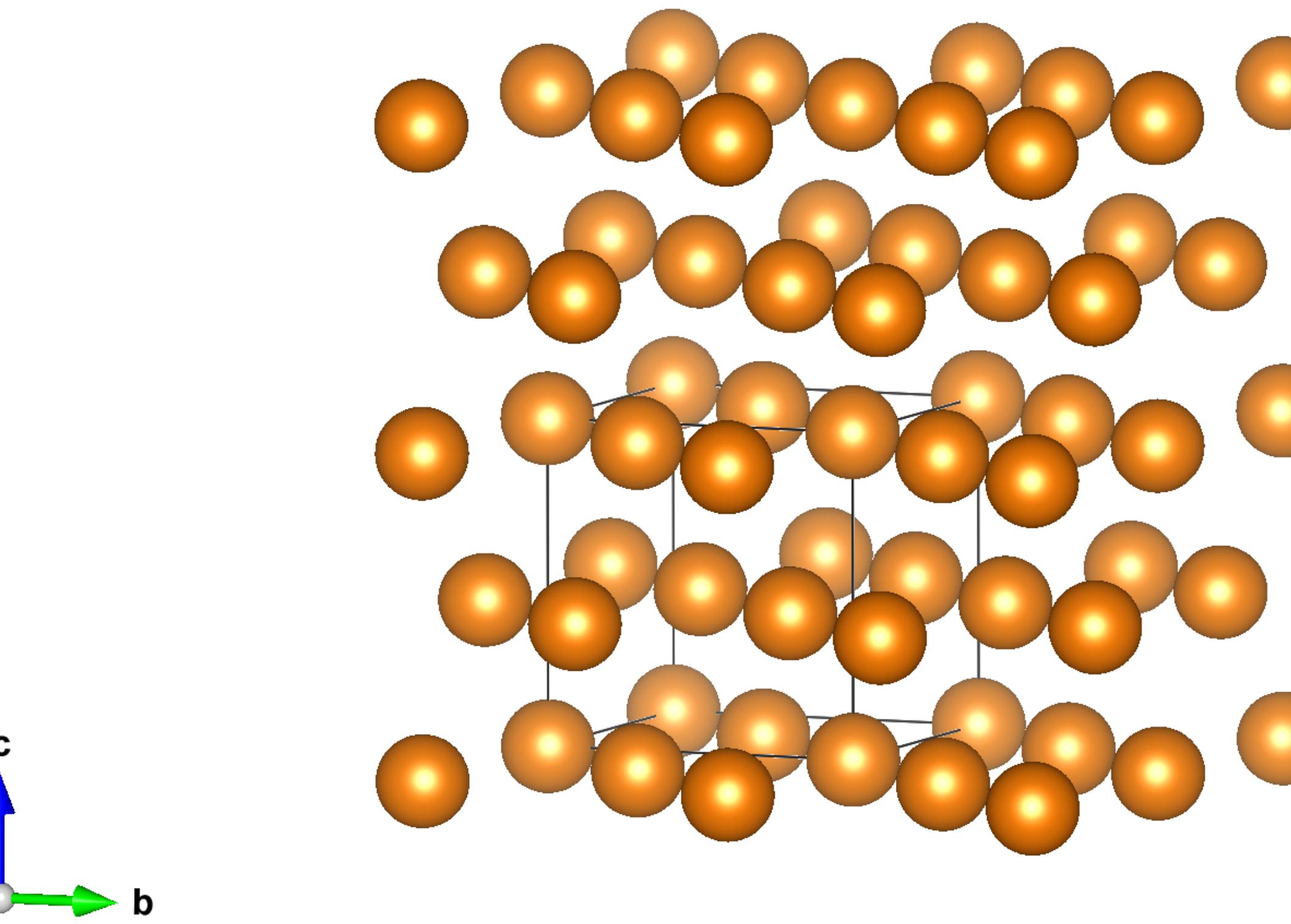
Inorganics
by space group

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
			(0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +
192	l	1	(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z) (z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y) (y,z,x) (-y,z,-x) (y,-z,x) (-y,-z,x) (y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z) (x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y) (z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x) (-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z) (-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y) (-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x) (-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z) (-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y) (-z,-y,x) (-z,y,-x) (z,-y,x) (z,y,x)
96	k	.m	(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z) (z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x) (x,z,x) (-x,z,-x) (x,-z,x) (-x,-z,x) (x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z) (x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x) (z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)
96	j	m..	(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z) (z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y) (y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0) (y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z) (0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y) (z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)
48	i	m.m 2	(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y) (y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y) (y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)
48	h	m.m 2	(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y) (y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y) (y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)
48	g	2.m m	(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4) (1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4) (x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)
32	f	.3m	(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x) (x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)
24	e	4m. m	(x,0,0) (-x,0,0) (0,x,0) (0,-x,0) (0,0,x) (0,0,-x)
24	d	m.m m	(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4) (1/4,1/4,0) (3/4,1/4,0)
8	c	-43m	(1/4,1/4,1/4) (1/4,1/4,3/4)
4	b	m-3m	(1/2,1/2,1/2)
4	a	m-3m	(0,0,0)

Wyckoff Positions of Group *Fm-3m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
			$(0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +$
192	l	1	$(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z)$ $(z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y)$ $(y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x)$ $(y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z)$ $(x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y)$ $(z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x)$ $(-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z)$ $(-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y)$ $(-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x)$ $(-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z)$ $(-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y)$ $(-z,-y,x) (-z,y,-x) (z,-y,x) (z,y,x)$
96	k	..m	$(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z)$ $(z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x)$ $(x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x)$ $(x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z)$ $(x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x)$ $(z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)$
96	j	m..	$(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z)$ $(z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y)$ $(y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0)$ $(y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z)$ $(0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y)$ $(z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)$
48	i	m.m 2	$(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y)$ $(y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y)$ $(y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)$
48	h	m.m 2	$(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y)$ $(y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y)$ $(y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)$
48	g	2.m m	$(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4)$ $(1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4)$ $(x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)$
32	f	.3m	$(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x)$ $(x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)$
24	e	4m. m	$(x,0,0) (-x,0,0) (0,x,0) (0,-x,0)$ $(0,0,x) (0,0,-x)$
24	d	m.m m	$(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4)$ $(1/4,1/4,0) (3/4,1/4,0)$
8	c	-43m	$(1/4,1/4,1/4) (1/4,1/4,3/4)$
4	b	m-3m	$(1/2,1/2,1/2)$
4	a	m-3m	$(0,0,0)$

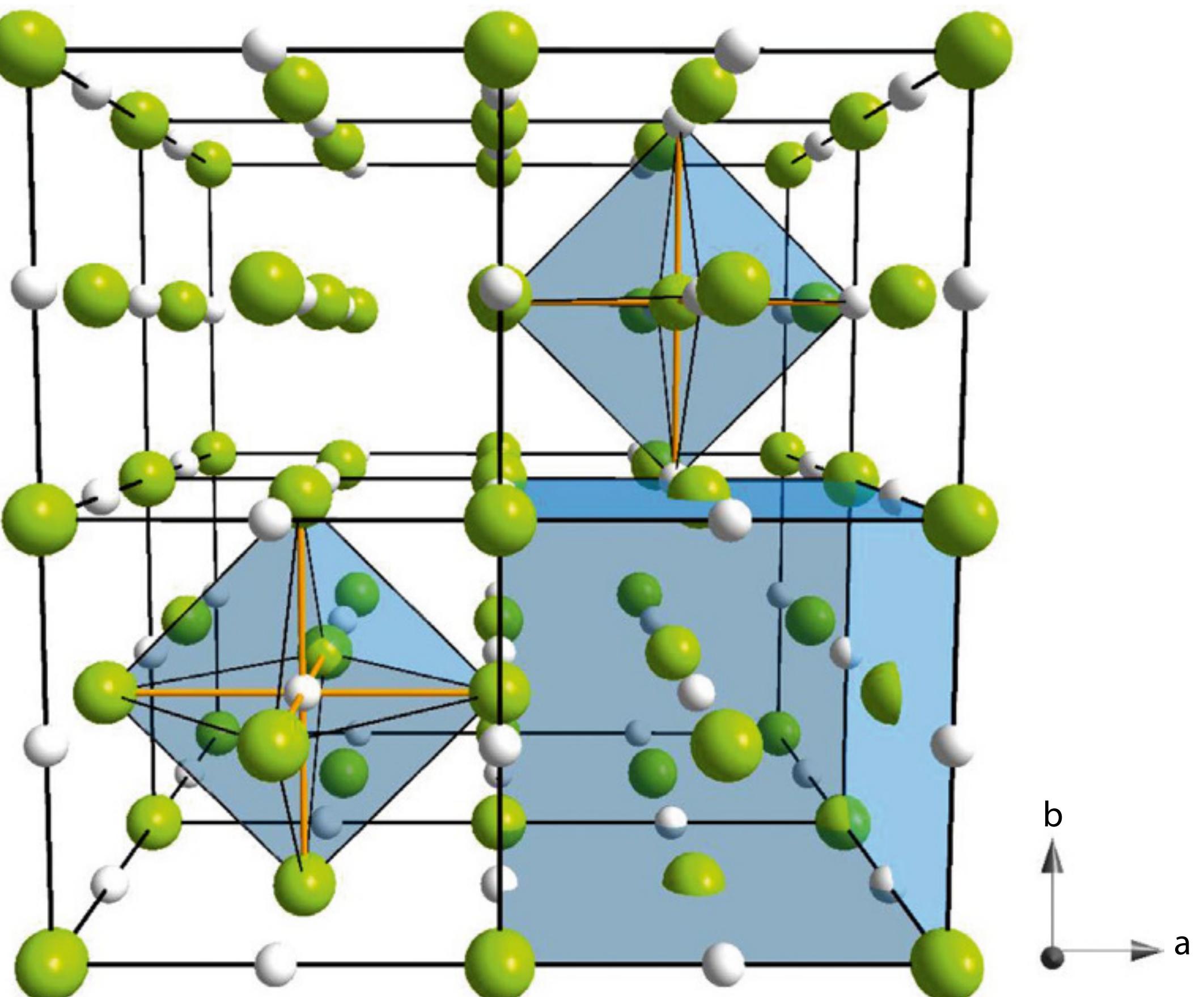
Copper



Wyckoff Positions of Group *Fm-3m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
			$(0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +$
192	l	1	$(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z)$ $(z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y)$ $(y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x)$ $(y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z)$ $(x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y)$ $(z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x)$ $(-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z)$ $(-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y)$ $(-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x)$ $(-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z)$ $(-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y)$ $(-z,-y,x) (-z,y,-x) (z,-y,x) (z,y,x)$
96	k	..m	$(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z)$ $(z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x)$ $(x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x)$ $(x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z)$ $(x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x)$ $(z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)$
96	j	m..	$(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z)$ $(z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y)$ $(y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0)$ $(y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z)$ $(0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y)$ $(z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)$
48	i	m.m 2	$(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y)$ $(y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y)$ $(y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)$
48	h	m.m 2	$(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y)$ $(y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y)$ $(y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)$
48	g	2.m m	$(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4)$ $(1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4)$ $(x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)$
32	f	.3m	$(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x)$ $(x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)$
24	e	4m. m	$(x,0,0) (-x,0,0) (0,x,0) (0,-x,0)$ $(0,0,x) (0,0,-x)$
24	d	m.m m	$(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4)$ $(1/4,1/4,0) (3/4,1/4,0)$
8	c	-43m	$(1/4,1/4,1/4) (1/4,1/4,3/4)$
4	b	m-3m	$(1/2,1/2,1/2)$
4	a	m-3m	$(0,0,0)$

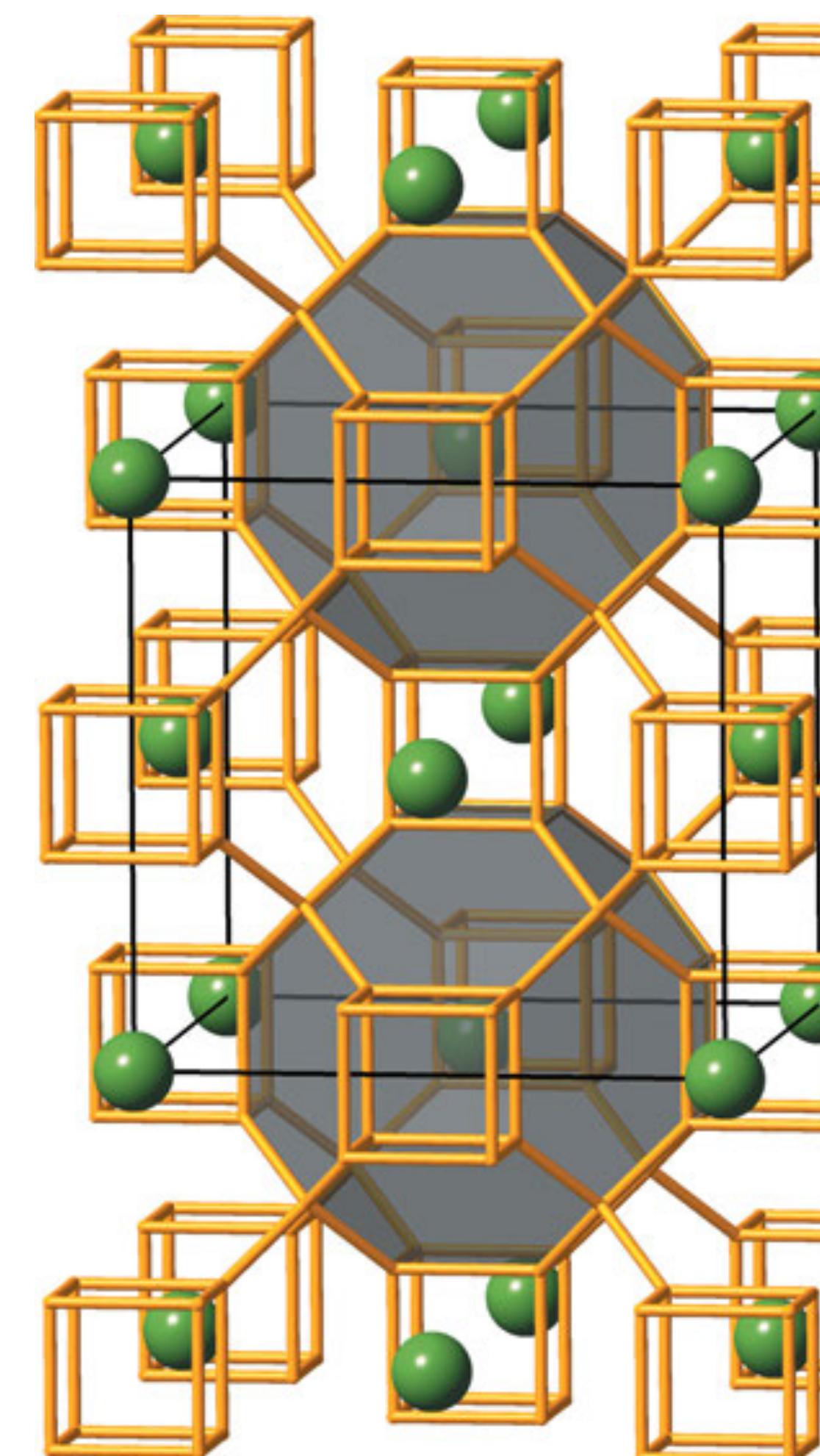
NaCl



Wyckoff Positions of Group *Fm-3m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
			$(0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +$
192	l	1	$(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z)$ $(z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y)$ $(y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x)$ $(y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z)$ $(x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y)$ $(z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x)$ $(-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z)$ $(-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y)$ $(-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x)$ $(-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z)$ $(-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y)$ $(-z,-y,x) (-z,y,-x) (z,-y,x) (z,y,x)$
96	k	..m	$(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z)$ $(z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x)$ $(x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x)$ $(x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z)$ $(x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x)$ $(z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)$
96	j	m..	$(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z)$ $(z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y)$ $(y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0)$ $(y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z)$ $(0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y)$ $(z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)$
48	i	m.m 2	$(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y)$ $(y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y)$ $(y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)$
48	h	m.m 2	$(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y)$ $(y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y)$ $(y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)$
48	g	2.m m	$(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4)$ $(1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4)$ $(x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)$
32	f	.3m	$(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x)$ $(x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)$
24	e	4m. m	$(x,0,0) (-x,0,0) (0,x,0) (0,-x,0)$ $(0,0,x) (0,0,-x)$
24	d	m.m m	$(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4)$ $(1/4,1/4,0) (3/4,1/4,0)$
8	c	-43m	$(1/4,1/4,1/4) (1/4,1/4,3/4)$
4	b	m-3m	$(1/2,1/2,1/2)$
4	a	m-3m	$(0,0,0)$

LaH₁₀

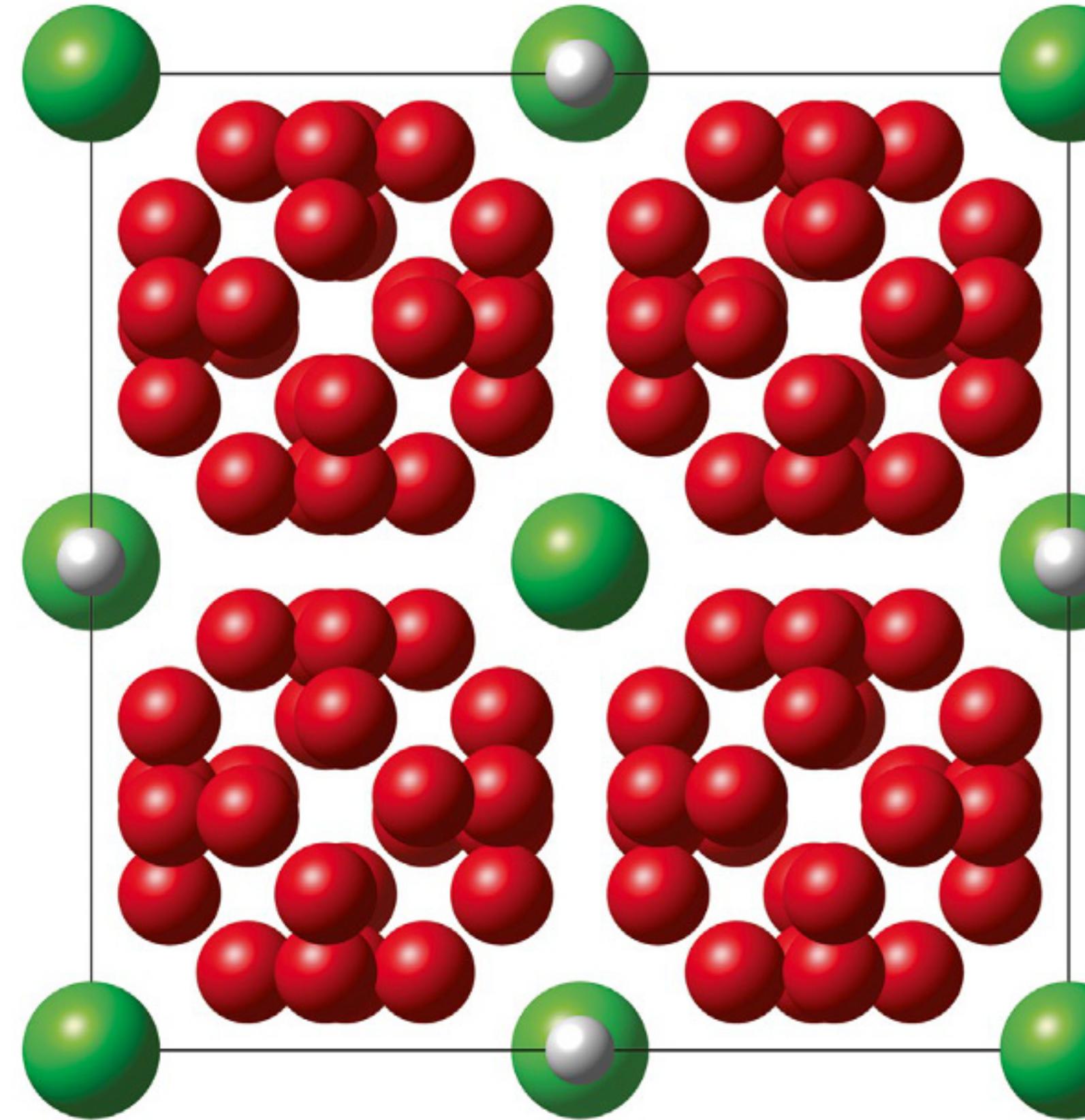


Wyckoff Positions of Group *Fm-3m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
			$(0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +$
192	l	1	$(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z)$ $(z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y)$ $(y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x)$ $(y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z)$ $(x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y)$ $(z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x)$ $(-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z)$ $(-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y)$ $(-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x)$ $(-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z)$ $(-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y)$ $(-z,-y,x) (-z,y,-x) (z,-y,x) (z,y,x)$
96	k	..m	$(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z)$ $(z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x)$ $(x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x)$ $(x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z)$ $(x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x)$ $(z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)$
96	j	m..	$(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z)$ $(z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y)$ $(y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0)$ $(y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z)$ $(0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y)$ $(z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)$
48	i	m.m 2	$(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y)$ $(y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y)$ $(y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)$
48	h	m.m 2	$(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y)$ $(y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y)$ $(y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)$
48	g	2.m m	$(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4)$ $(1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4)$ $(x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)$
32	f	.3m	$(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x)$ $(x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)$
24	e	4m. m	$(x,0,0) (-x,0,0) (0,x,0) (0,-x,0)$ $(0,0,x) (0,0,-x)$
24	d	m.m m	$(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4)$ $(1/4,1/4,0) (3/4,1/4,0)$
8	c	-43m	$(1/4,1/4,1/4) (1/4,1/4,3/4)$
4	b	m-3m	$(1/2,1/2,1/2)$
4	a	m-3m	$(0,0,0)$

NaCl “X₄₈”

Frank Hoffmann, Introduction to Crystallography (2020)



The crystal space is big
but we only care about a tiny corner of it

Crystal generative model

$$p(\mathbf{C} | g)$$

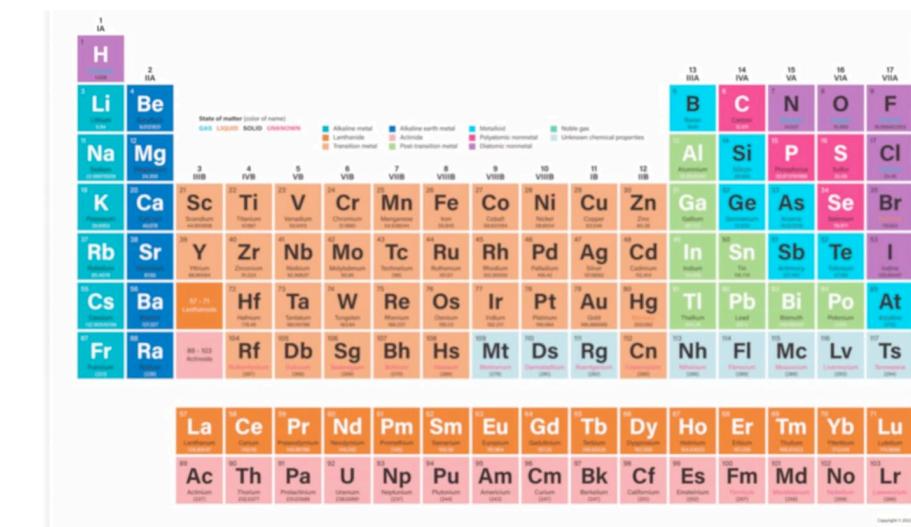
W

Wyckoff letters

a,b,c,d,e,...

A

Chemical species



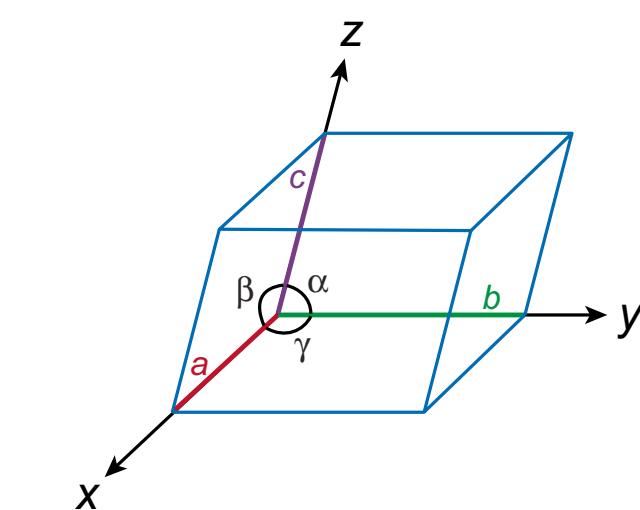
X

Fractional coordinates

(x_i, y_i, z_i)

L

Lattice parameters



CaCO_3

i	w_i	a_i	x_i	y_i	z_i
1	a	C	0	0	$\frac{1}{4}$
2	b	Ca	0	0	0
3	e	O	0.257	0	$\frac{1}{4}$
4	X	-	-	-	-

W

Wyckoff letters

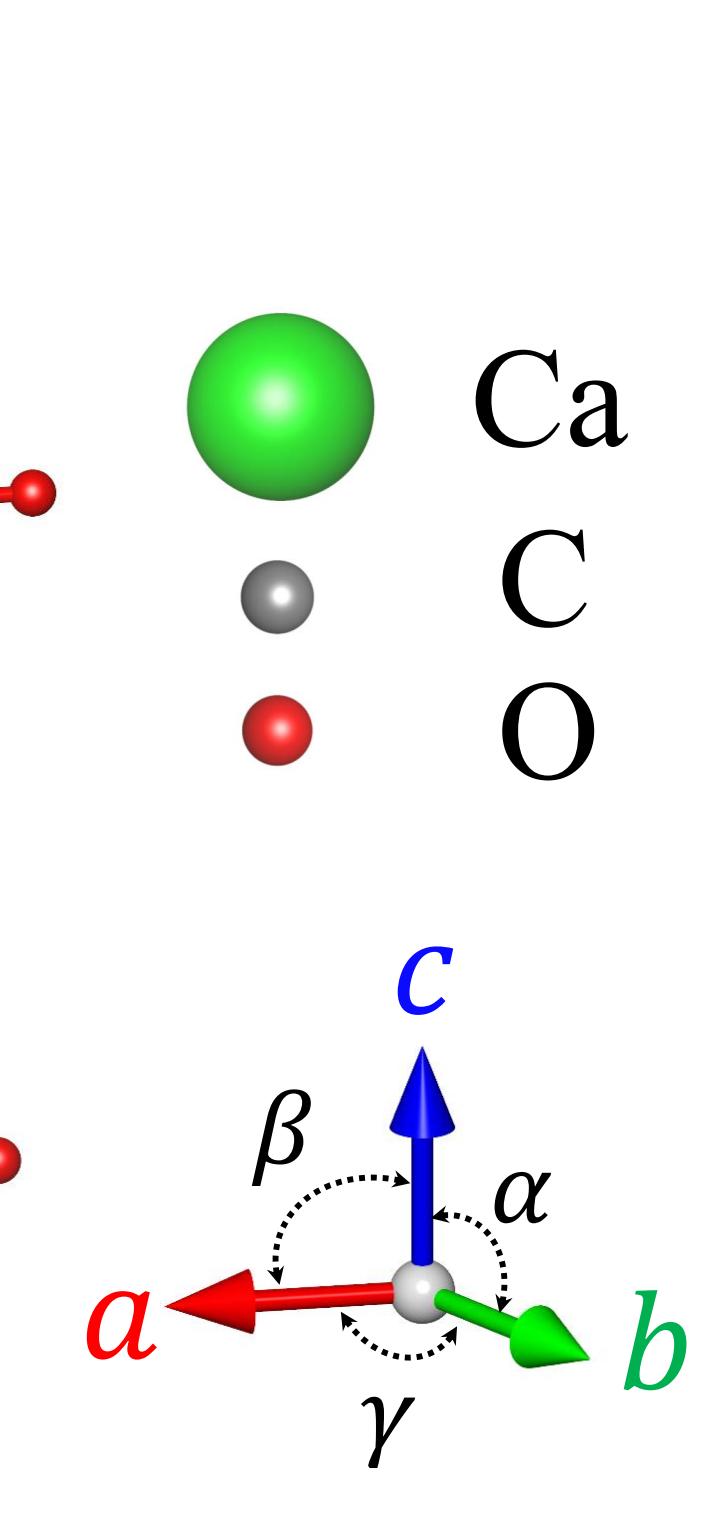
A

Chemical species

X

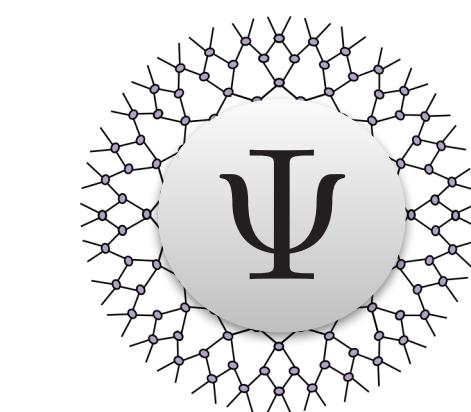
Fractional coordinates

L



Lattice parameters

Generative models and their physics genes

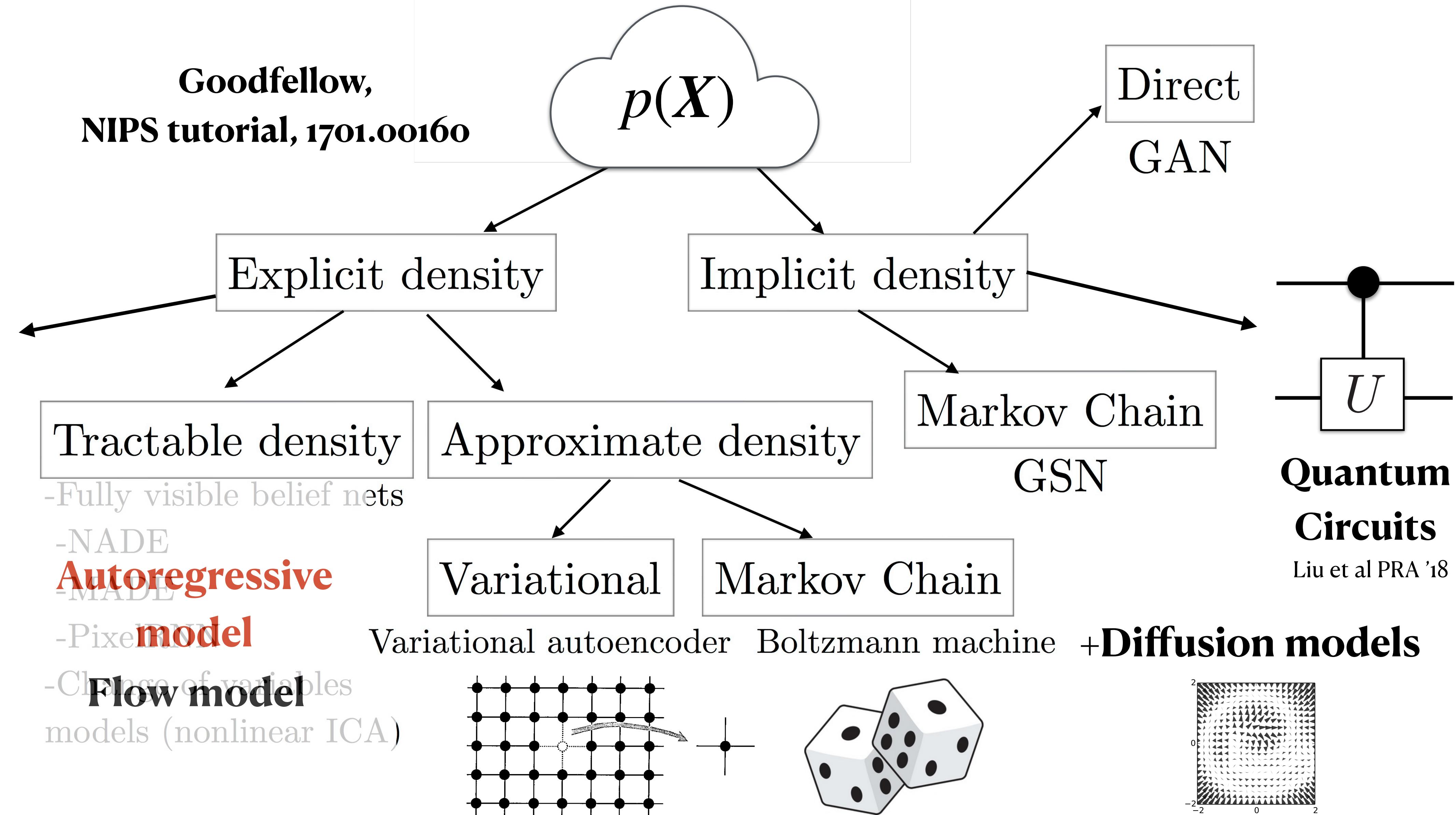


Tensor Networks

Han et al, PRX '18

Autoregressive
-NADE
-MADE
-PixelRNN
-Change of variables
Flow model
models (nonlinear ICA)

Goodfellow,
NIPS tutorial, 1701.00160

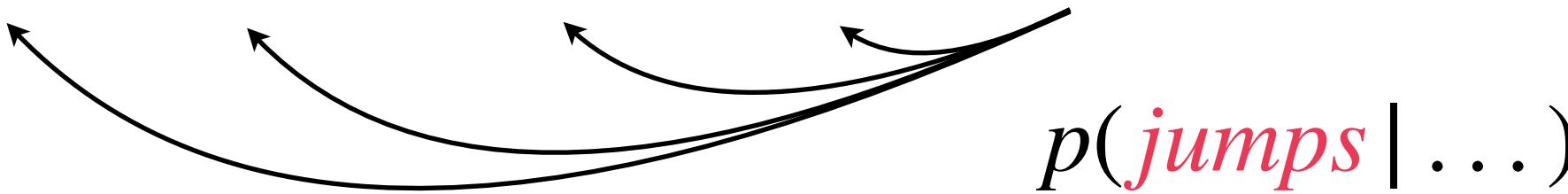


Autoregressive models

$$p(X) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2)\cdots$$

Language: GPT 2005.14165

“... *quick brown fox jumps* ...”



$p(\text{jumps} | \dots)$

Speech: WaveNet 1609.03499

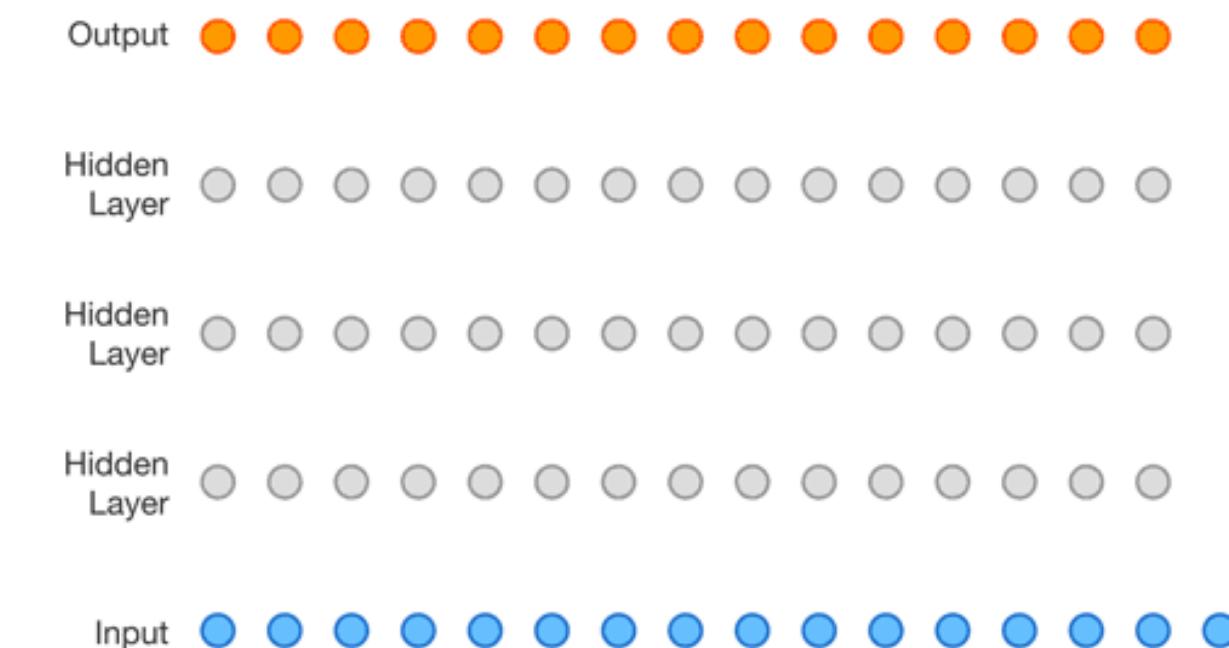
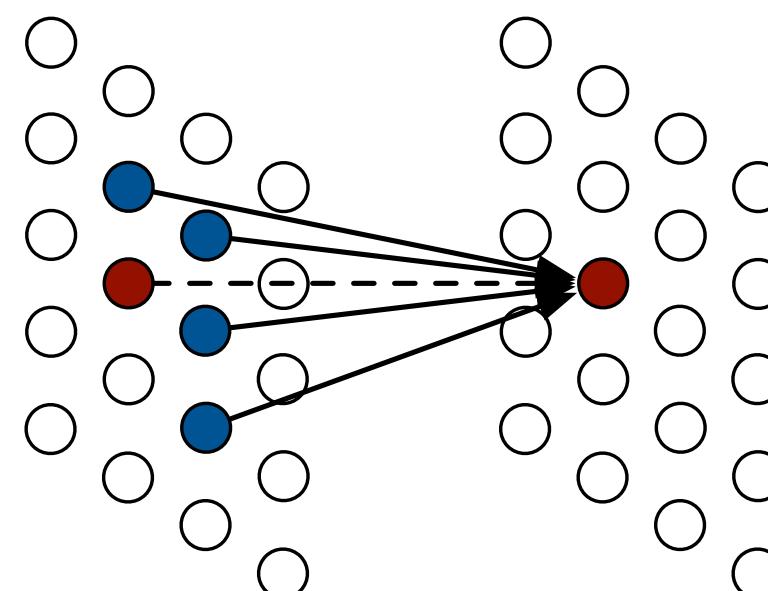
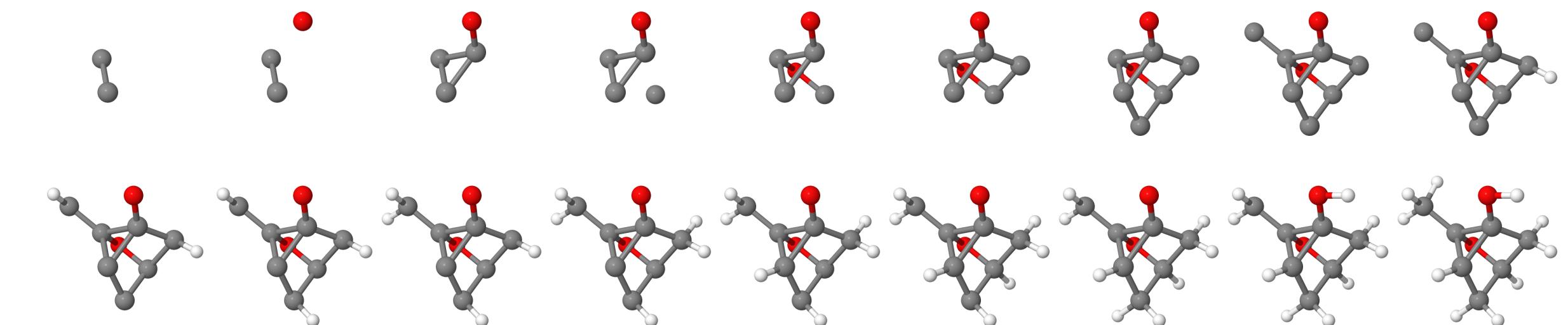


Image: PixelCNN 1601.06759



Molecular graph: 1810.11347



Autoregressive models

$$p(X) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)\cdots$$

Language: GPT 2005.14165

Speech: WaveNet 1609.03499

“... *quick brown fox jumps* ...”



$p(\text{jumps} \mid \dots)$

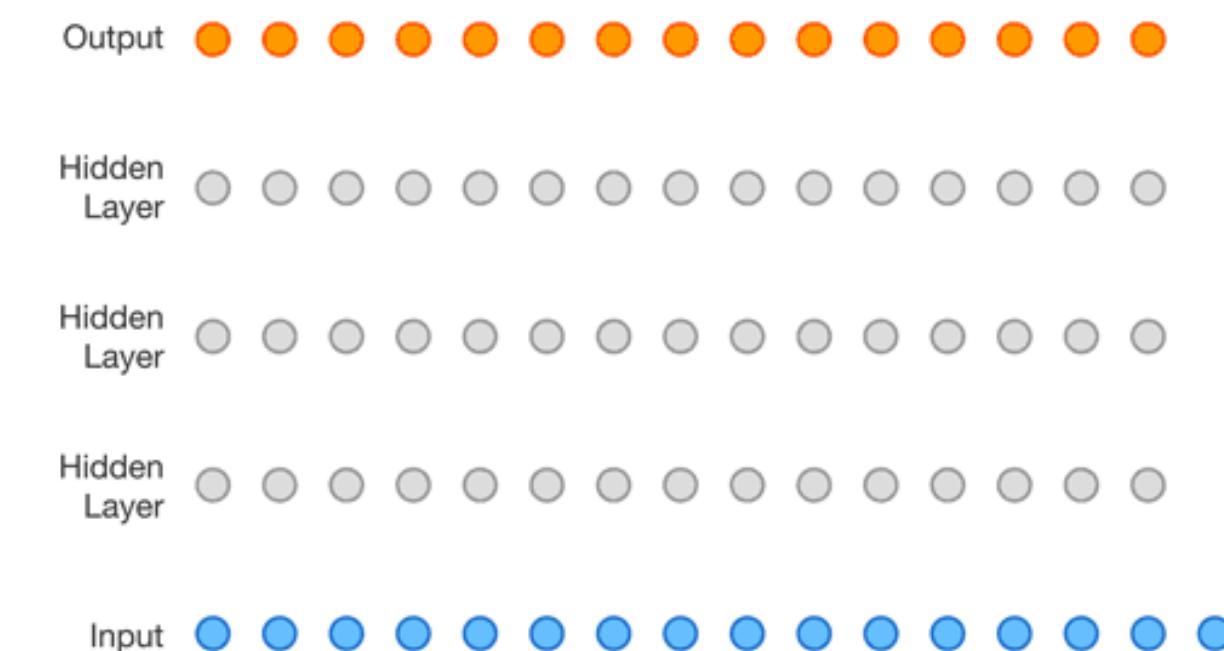
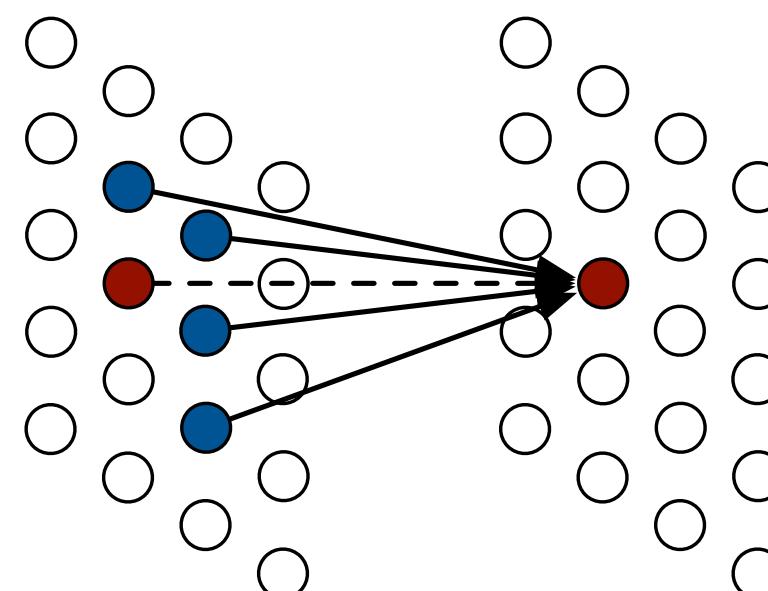
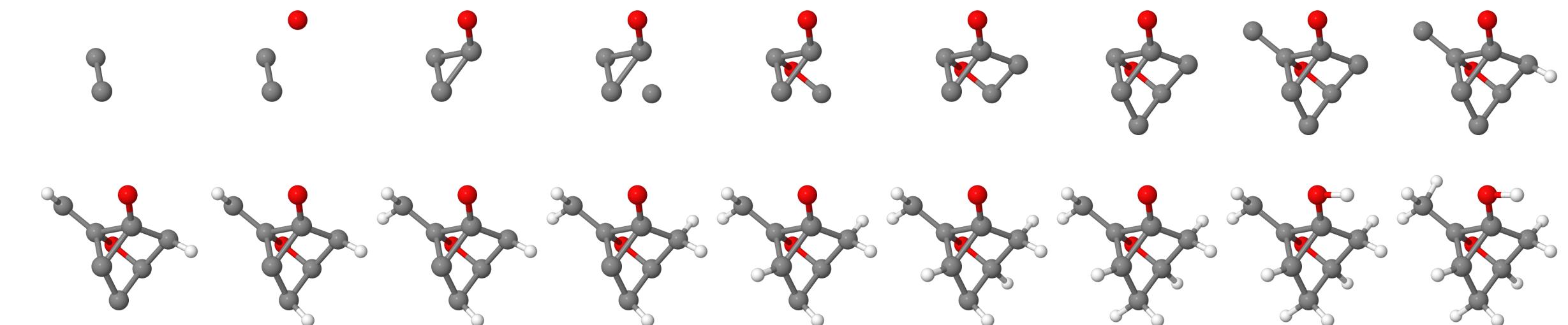


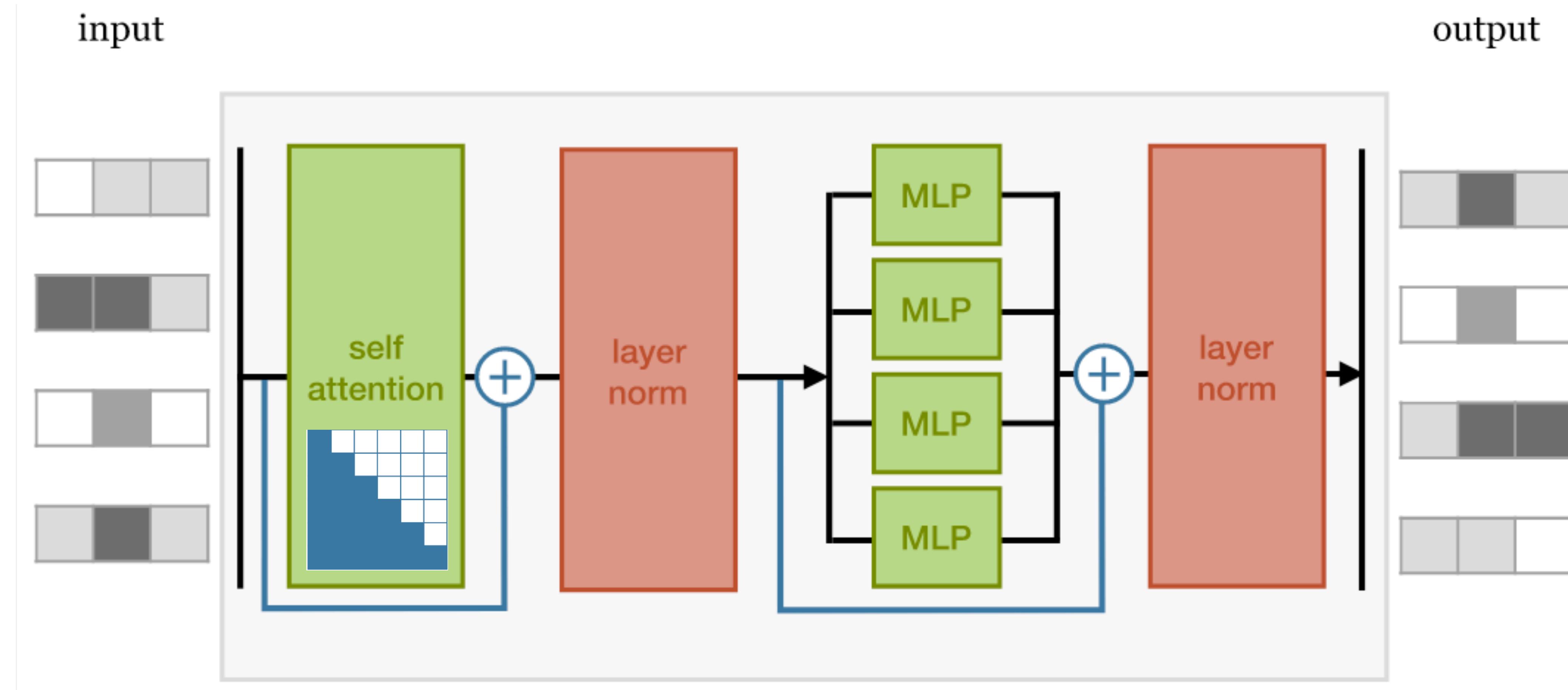
Image: PixelCNN 1601.06759



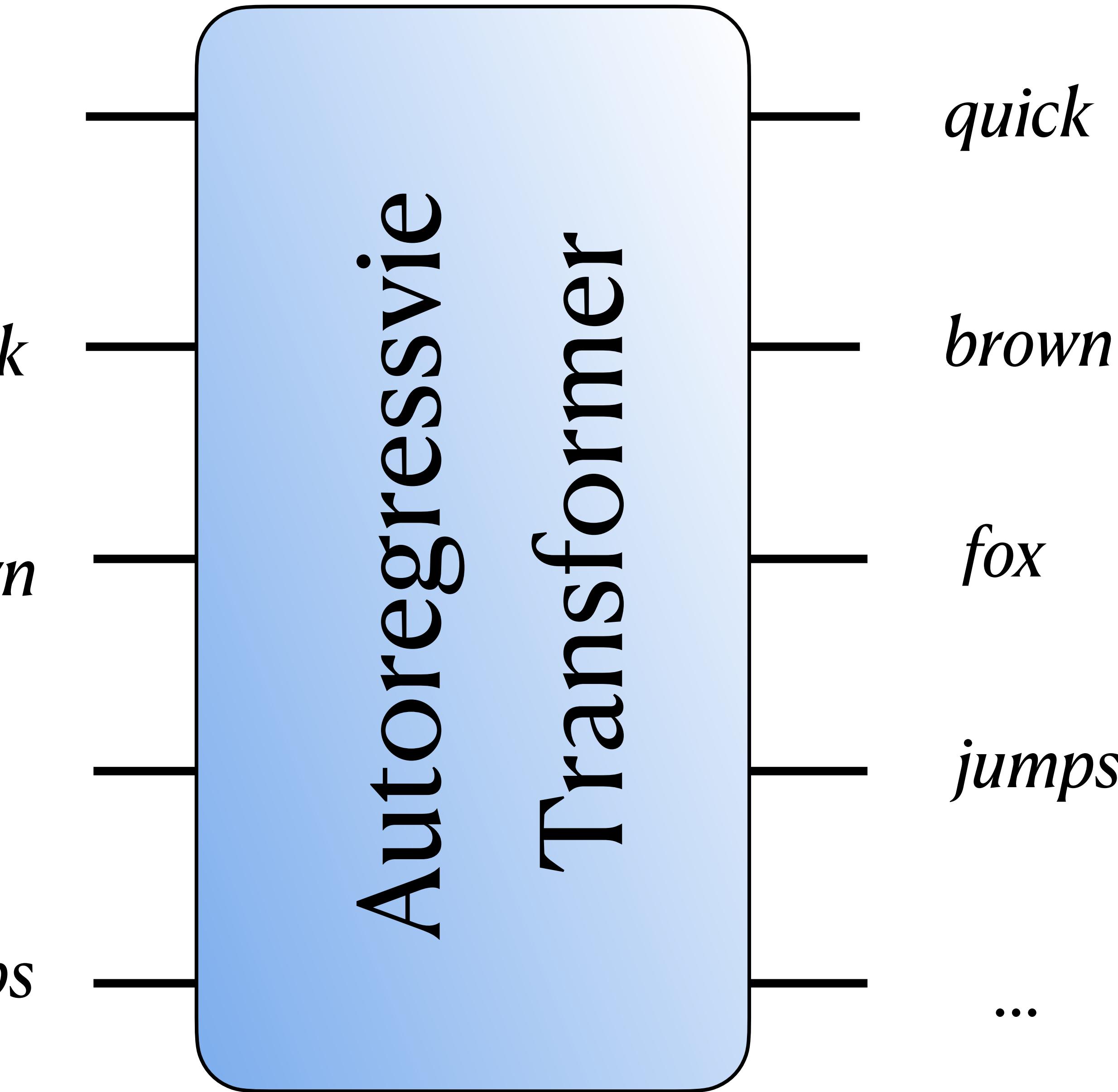
Molecular graph: 1810.11347



The transformer



Great at capturing long-range dependence; friendly to backpropagation and GPUs
Masked attention matrix => lower triangular Jacobian matrix => autoregressive model



Language = anything you can tokenize and serialize

Autoregressive crystal model

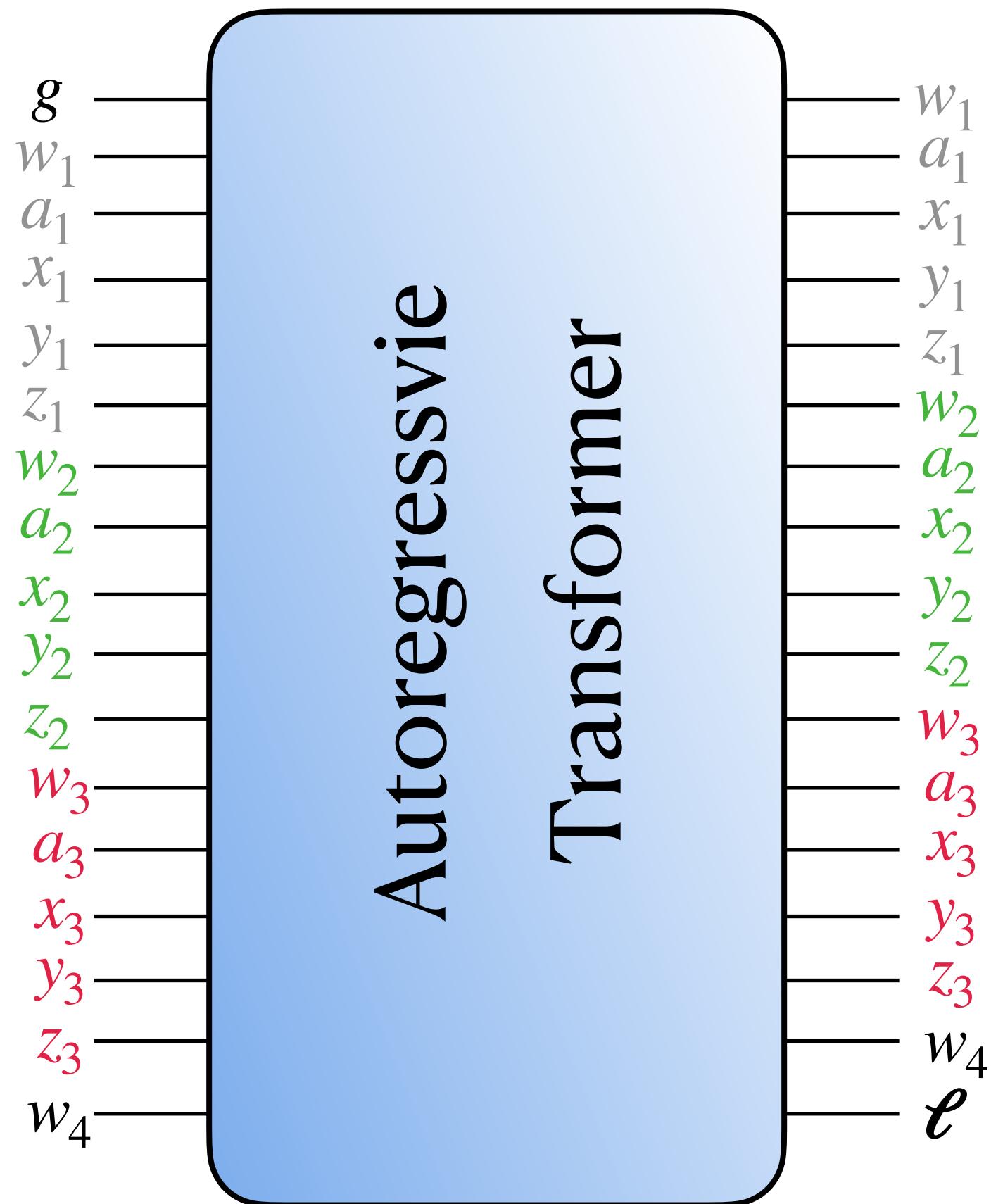
i	w_i	a_i	x_i	y_i	z_i
1	a	C	0	0	$\frac{1}{4}$
2	b	Ca	0	0	0
3	e	O	0.257	0	$\frac{1}{4}$
4	X	-	-	-	-

Autoregressive crystal model

i	w_i	a_i	x_i	y_i	z_i
1	a	C	0	0	$\frac{1}{4}$
2	b	Ca	0	0	0
3	e	O	0.257	0	$\frac{1}{4}$
4	X	-	-	-	-

167-a-C-o-o-1/4-b-Ca-o-o-o-e-O-o-0.257-o-1/4-X-4.99Å-4.99Å-17.07Å-90°-90°-120°

CrystalFormer



g : 1-230 categorical

W : 'a'-'z' + 'A' + 'X'

A : 'H'-'Og'

X : continuous periodic variables

L : continuous variables

167-a-C-o-o-1/4-**b**-Ca-o-o-o-e-O-0.257-0-1/4-X-4.99-4.99-17.07-90-90-120

Grammar of the crystal language

167-a-C-o-o-1/4-b-Ca-o-o-o-e-O-o.257-o-1/4-X-4.99-4.99-17.07-90-90-120

We do not want to waste data and compute on these hard constraints

I worked on a problem with a similar constraint in the past (Variational density matrix for uniform electron gas: occupation in k-space with Pauli exclusion, Xie et al, 2201.03156, SciPost Physics '23).

Pure technical reason for kicking off this study. The other reason is GPT: 1) it proves that autoregressive transformers are powerful; 2) it teaches me crystallography and materials genomics libraries, writes 20% of the codes ...

Grammar of the crystal language

167-a-C-o-0-1/4-b-Ca-o-o-o-e-O-o.257-o-1/4-X-4.99-[REDACTED]-17.07-[REDACTED]

inactive lattice parameters

We do not want to waste data and compute on these hard constraints

I worked on a problem with a similar constraint in the past (Variational density matrix for uniform electron gas: occupation in k-space with Pauli exclusion, Xie et al, 2201.03156, SciPost Physics '23).

Pure technical reason for kicking off this study. The other reason is GPT: 1) it proves that autoregressive transformers are powerful; 2) it teaches me crystallography and materials genomics libraries, writes 20% of the codes ...

Grammar of the crystal language

inactive fractional coordinates

167-a-C-[REDACTED]-b-Ca-[REDACTED]-e-O-o.257-[REDACTED]-X-4.99-[REDACTED]-17.07-[REDACTED]

inactive lattice parameters

We do not want to waste data and compute on these hard constraints

I worked on a problem with a similar constraint in the past (Variational density matrix for uniform electron gas: occupation in k-space with Pauli exclusion, Xie et al, 2201.03156, SciPost Physics '23).

Pure technical reason for kicking off this study. The other reason is GPT: 1) it proves that autoregressive transformers are powerful; 2) it teaches me crystallography and materials genomics libraries, writes 20% of the codes ...

Grammar of the crystal language

inactive fractional coordinates

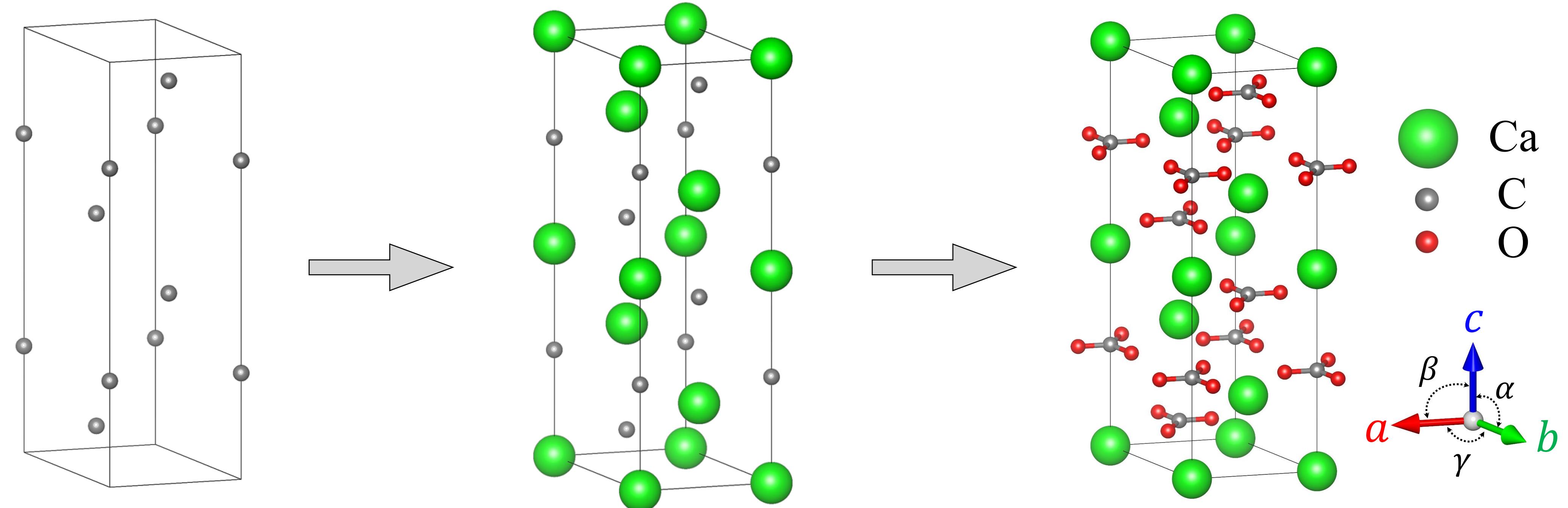


We do not want to waste data and compute on these hard constraints

I worked on a problem with a similar constraint in the past (Variational density matrix for uniform electron gas: occupation in k-space with Pauli exclusion, Xie et al, 2201.03156, SciPost Physics '23).

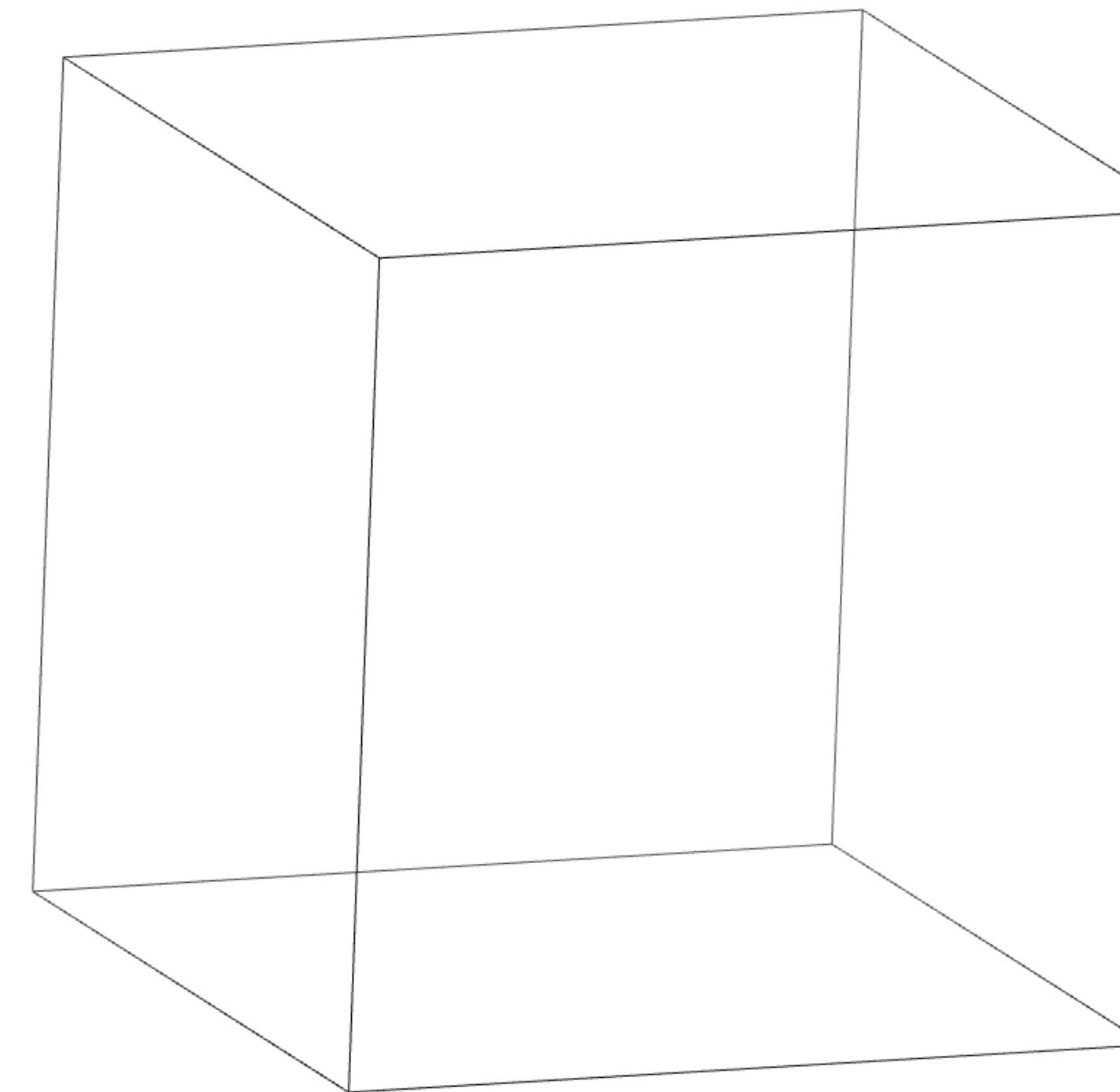
Pure technical reason for kicking off this study. The other reason is GPT: 1) it proves that autoregressive transformers are powerful; 2) it teaches me crystallography and materials genomics libraries, writes 20% of the codes ...

Autoregressive sampling of crystal

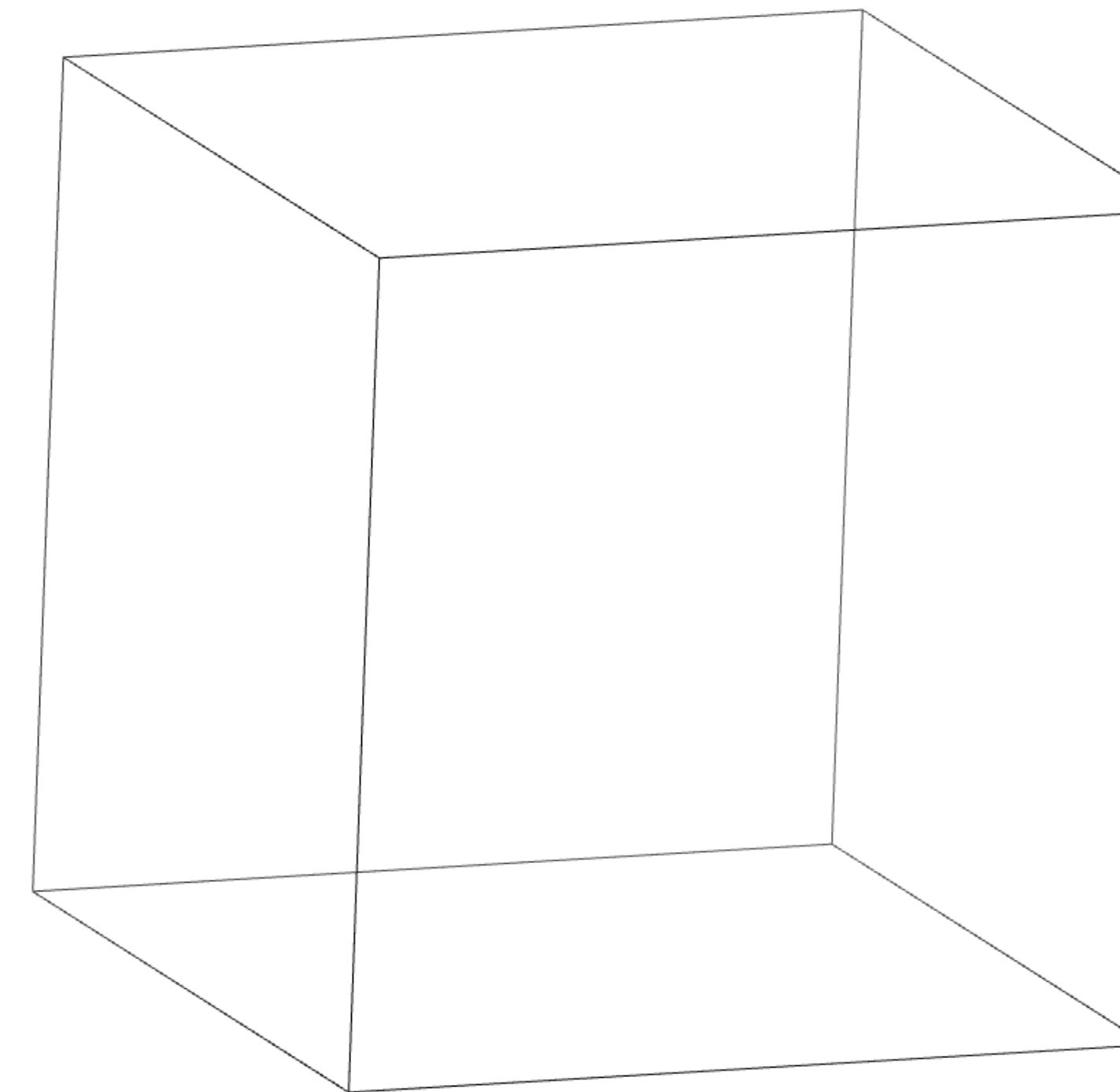


167-a-C-[REDACTED]-b-Ca-[REDACTED]-e-O-0.257-[REDACTED]-X-4.99-[REDACTED]-17.07-[REDACTED]

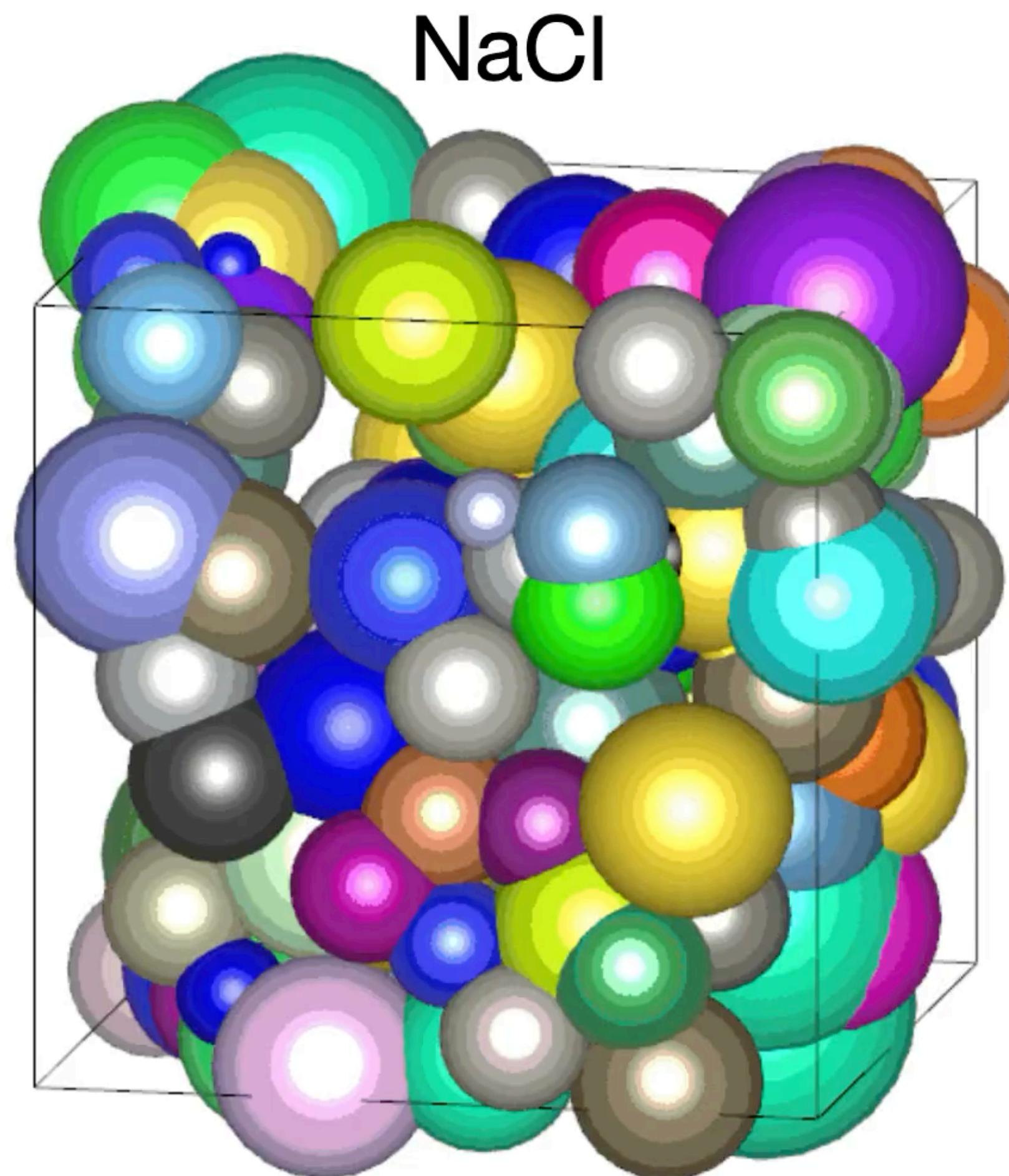
Autoregressive sampling of $\text{Cs}_2\text{ZnFe}(\text{CN})_6$



Autoregressive sampling of $\text{Cs}_2\text{ZnFe}(\text{CN})_6$



Generating crystals via diffusion



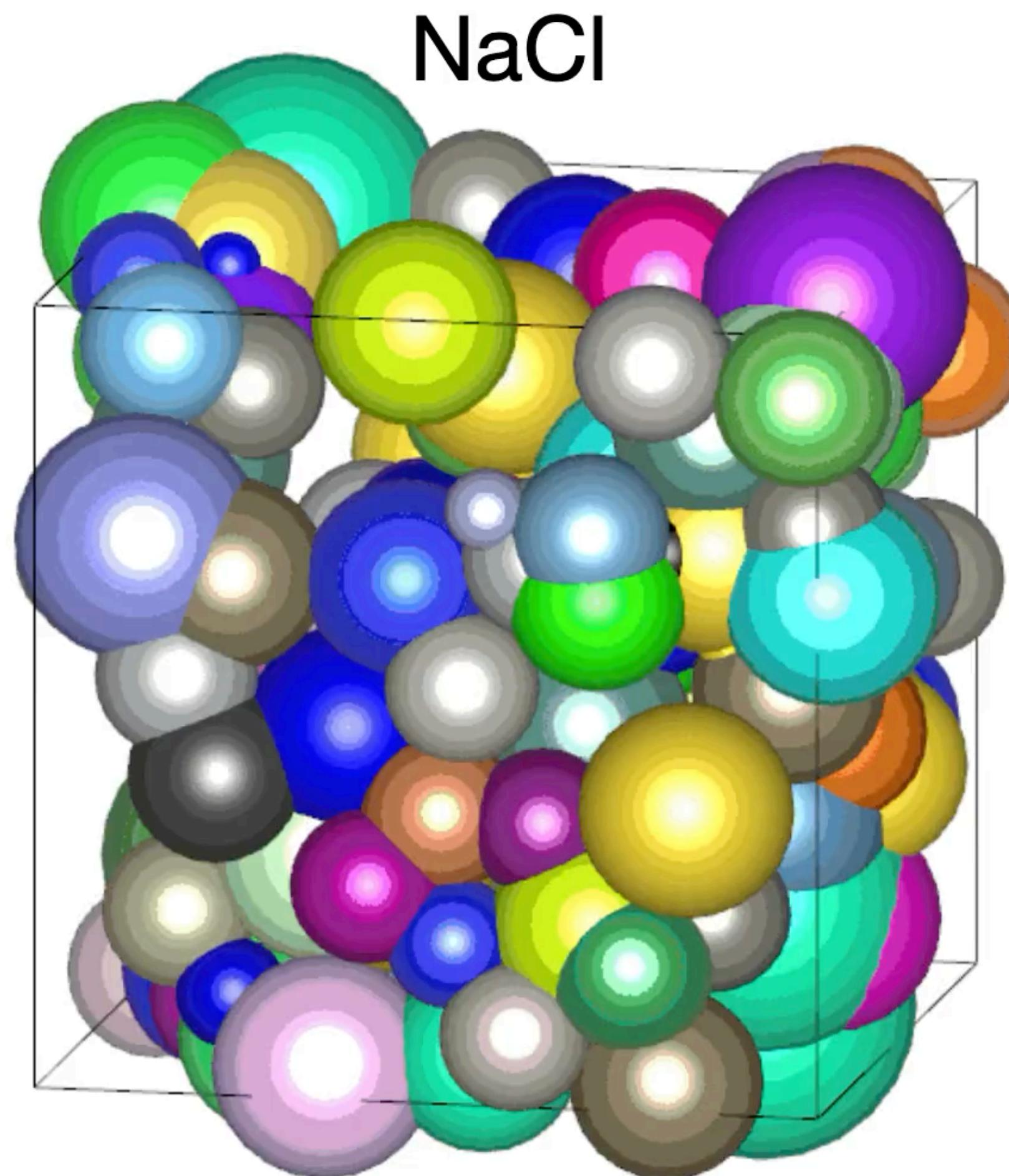
 Google DeepMind

Yang et al, 2311.09235

<https://unified-materials.github.io/unimat/>

See also Xie et al, 2110.06197, Zheng et al, 2306.05445, Jiao et al, 2309.04475, Zeni et al, 2312.03687, Jiao et al, 2402.03992 ...

Generating crystals via diffusion



 Google DeepMind

Yang et al, 2311.09235

<https://unified-materials.github.io/unimat/>

See also Xie et al, 2110.06197, Zheng et al, 2306.05445, Jiao et al, 2309.04475, Zeni et al, 2312.03687, Jiao et al, 2402.03992 ...

```
"""
data_CaCO3
.
.
.
_symmetry_space_group_name_H-M      R-3c
_cell_length_a    4.99
_cell_length_b    4.99
_cell_length_c    17.06
_cell_angle_alpha 90.0
_cell_angle_beta   90.00
_cell_angle_gamma 120.0
_symmetry_Int_Tables_number      167
_chemical_formula_structural     CaCO3
_chemical_formula_sum           'Ca6 C6 O18'
_cell_volume                  368.1
_cell_formula_units_Z          6
.
.
.
_atom_site_type_symbol
_atom_site_label
_atom_site_symmetry_multiplicity
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_occupancy
Ca  Ca0  6  0.000  0.000  0.000  1
C   C1   6  0.0000  0.000  0.250  1
O   O2   18  0.257 0.000  0.250  1
"""

```

The LLM learns to generate CIF as pure texts

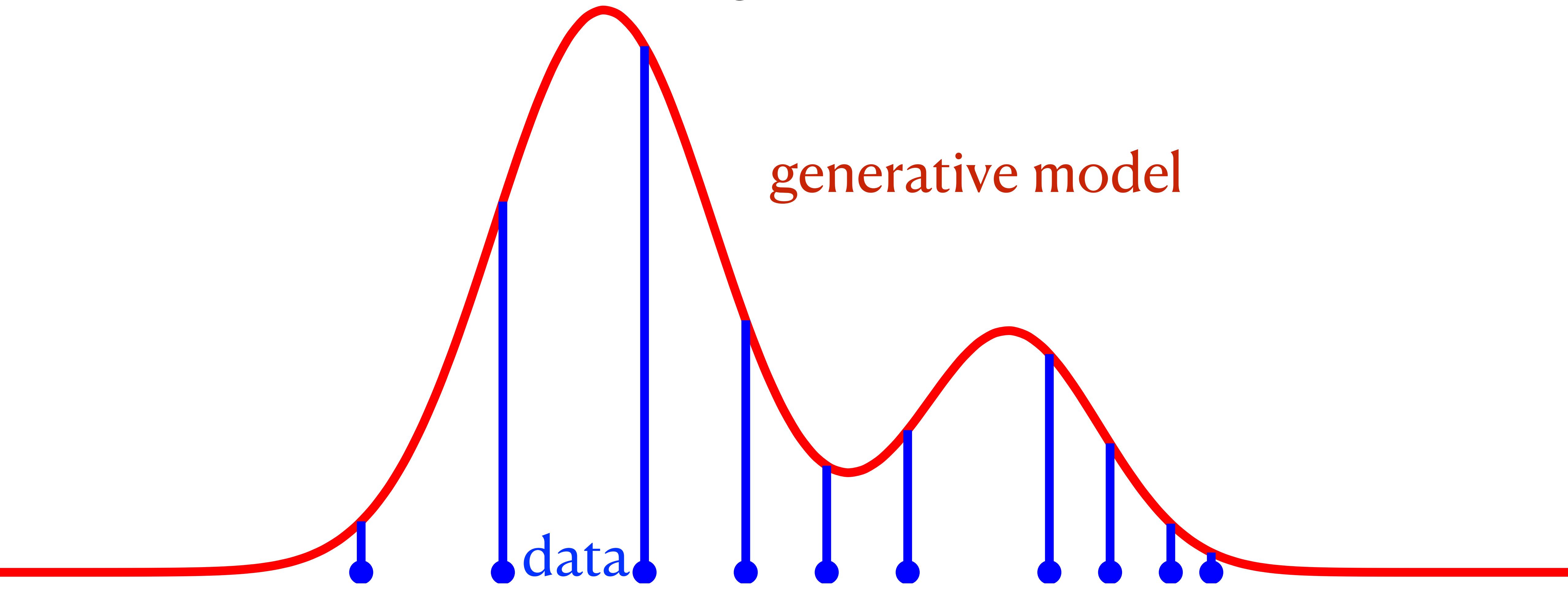
A lot more redundant information=>
large vocabulary, longer sequences,
slower sampling

One has to learn the space group—
Wyckoff position—multiplicities relation
from data

*CrystalFormer speaks more concise
and precise math language*

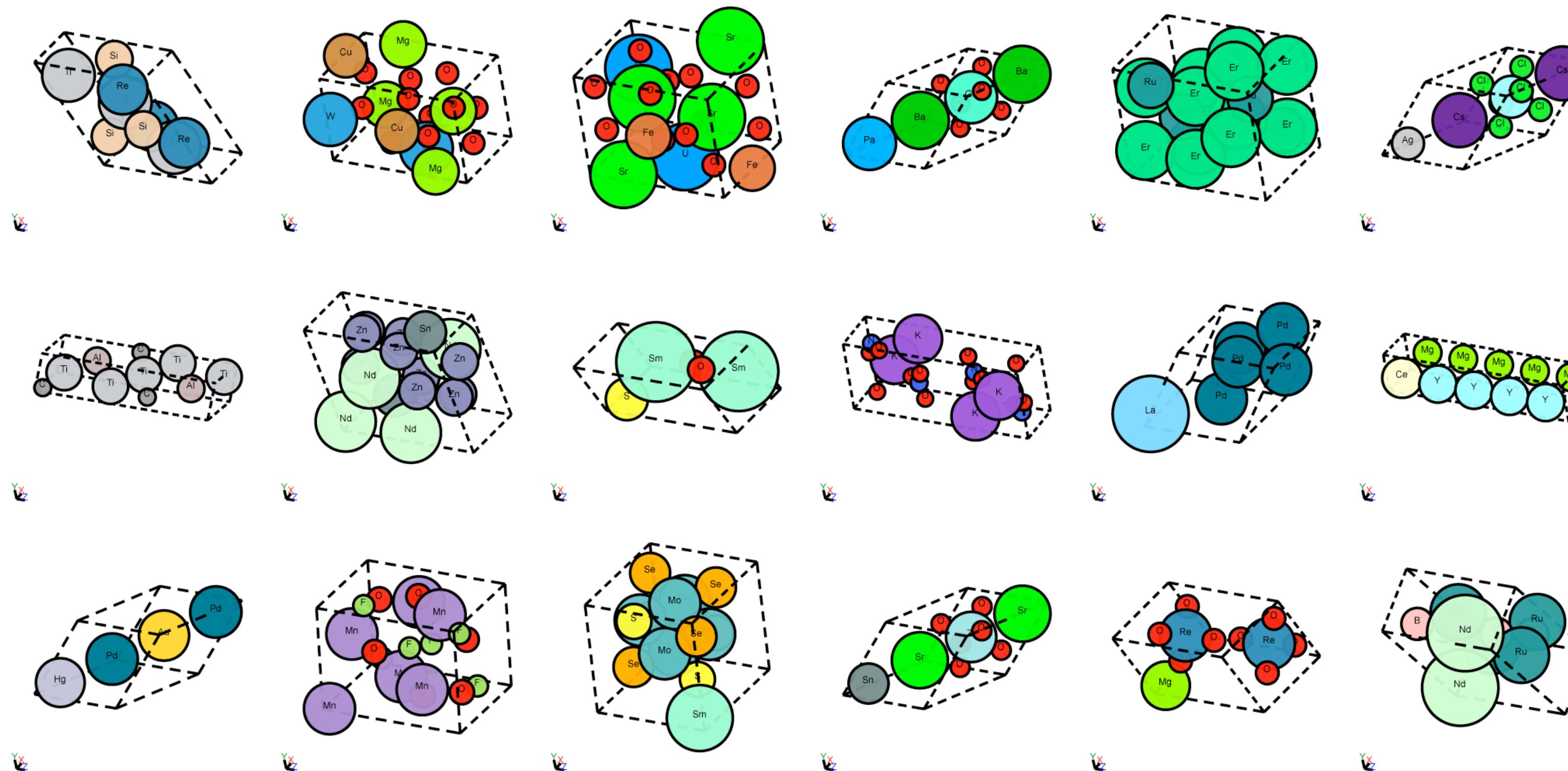
Training: maximum likelihood estimation

$$\mathcal{L} = - \mathbb{E}_{C,g} [\ln p(C|g)]$$

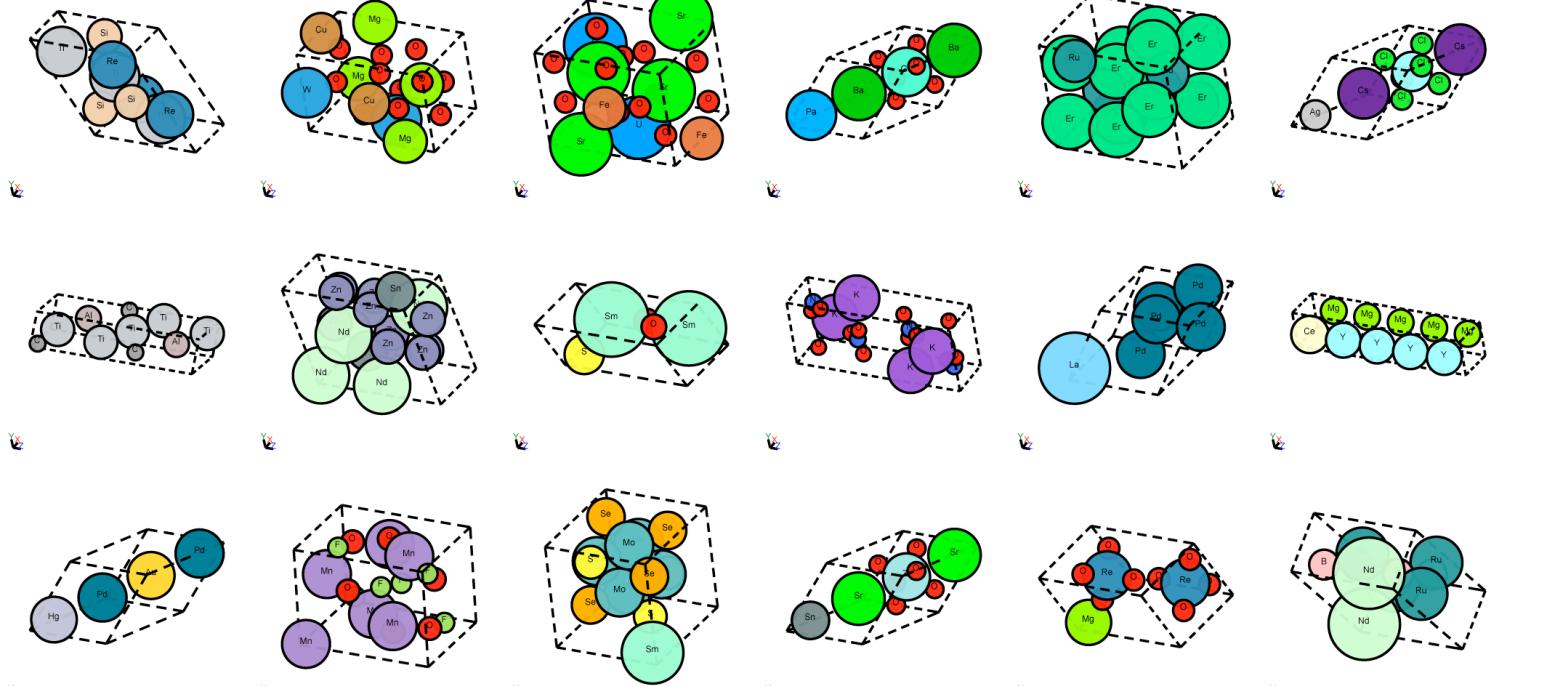


Training data

https://github.com/txie-93/cdvae/tree/main/data/mp_20

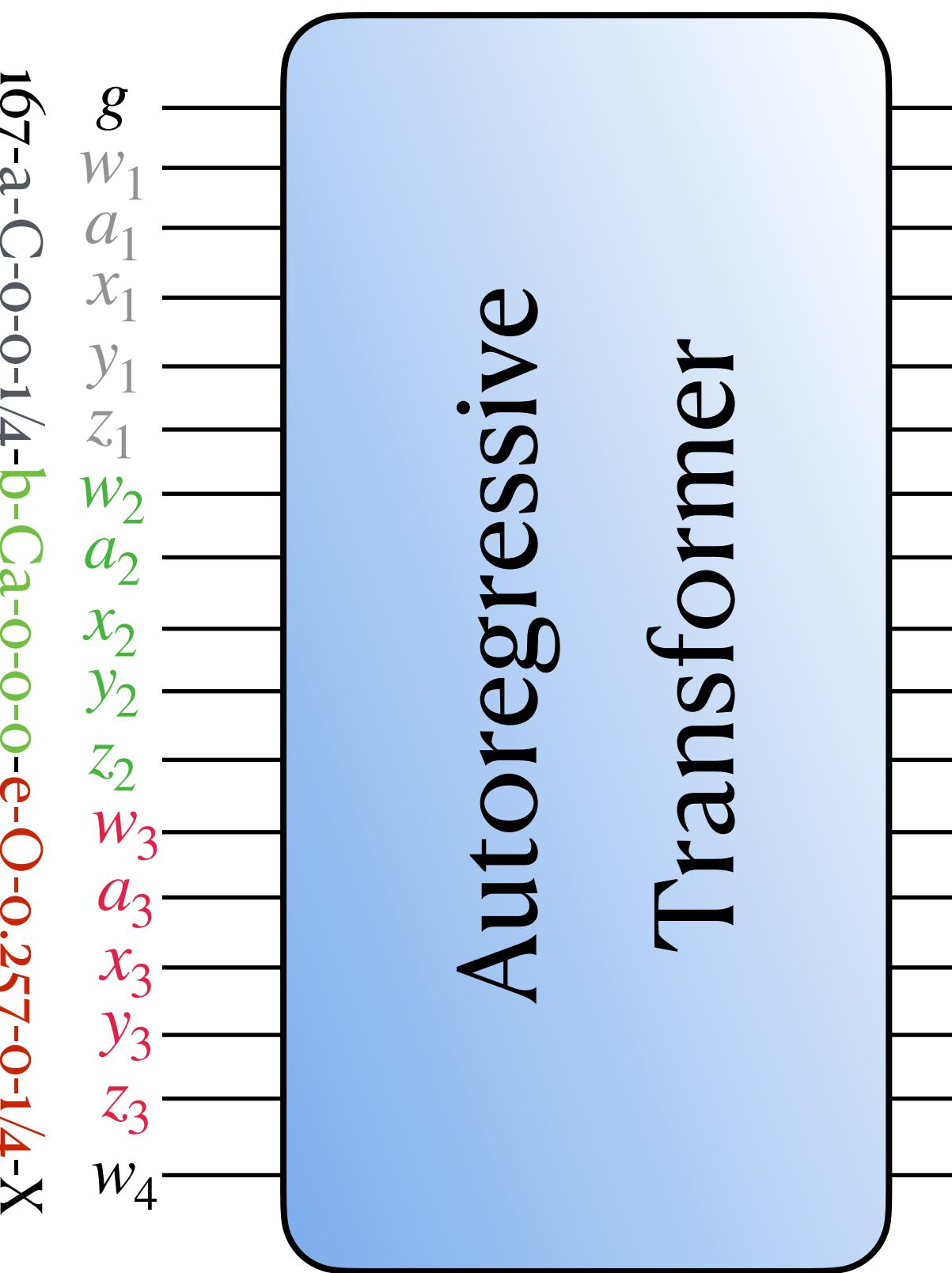


MP-20 contains 45231 inorganic materials which includes almost all experimentally stable materials from the Materials Project with unit cells including at most 20 atoms.

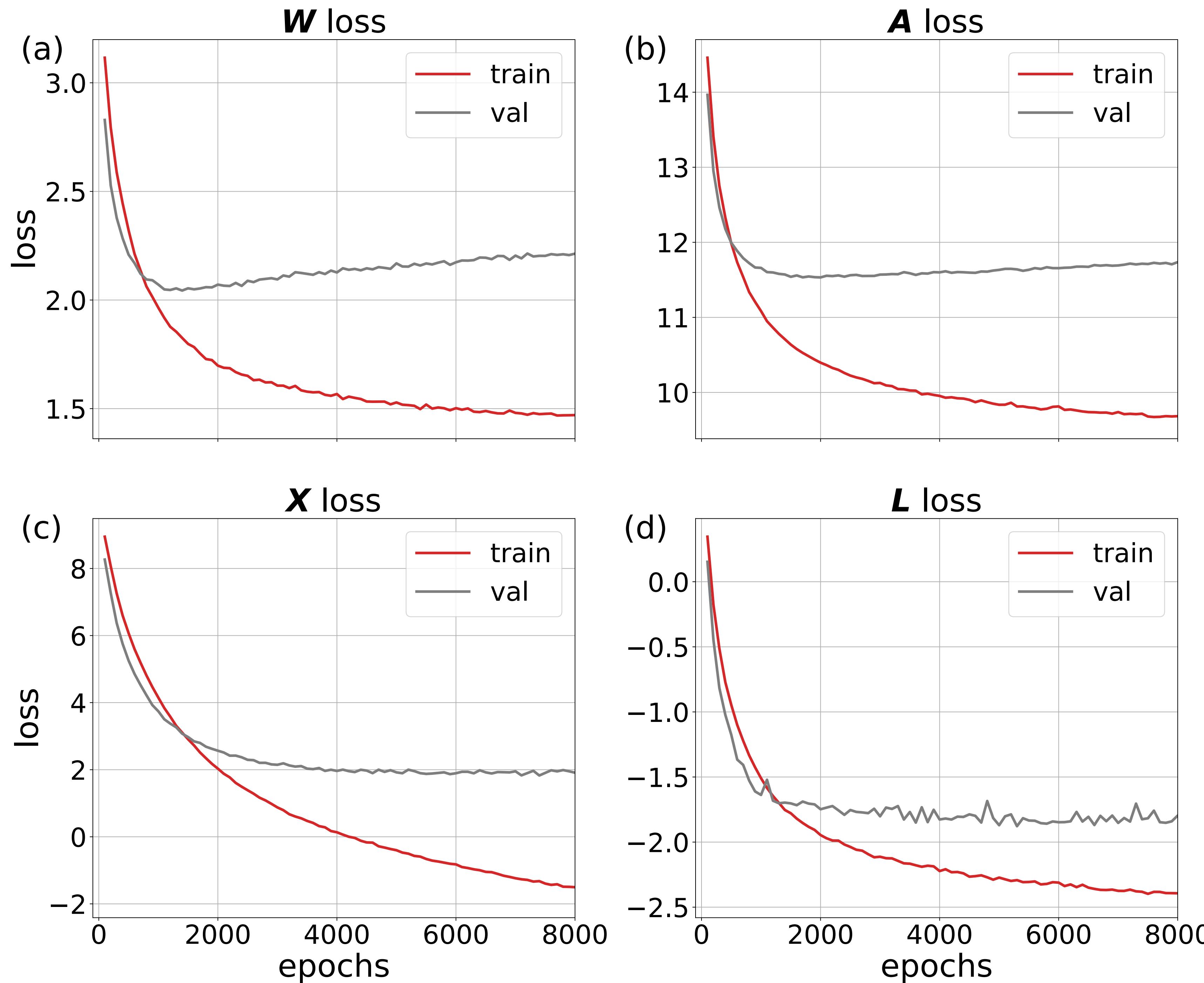


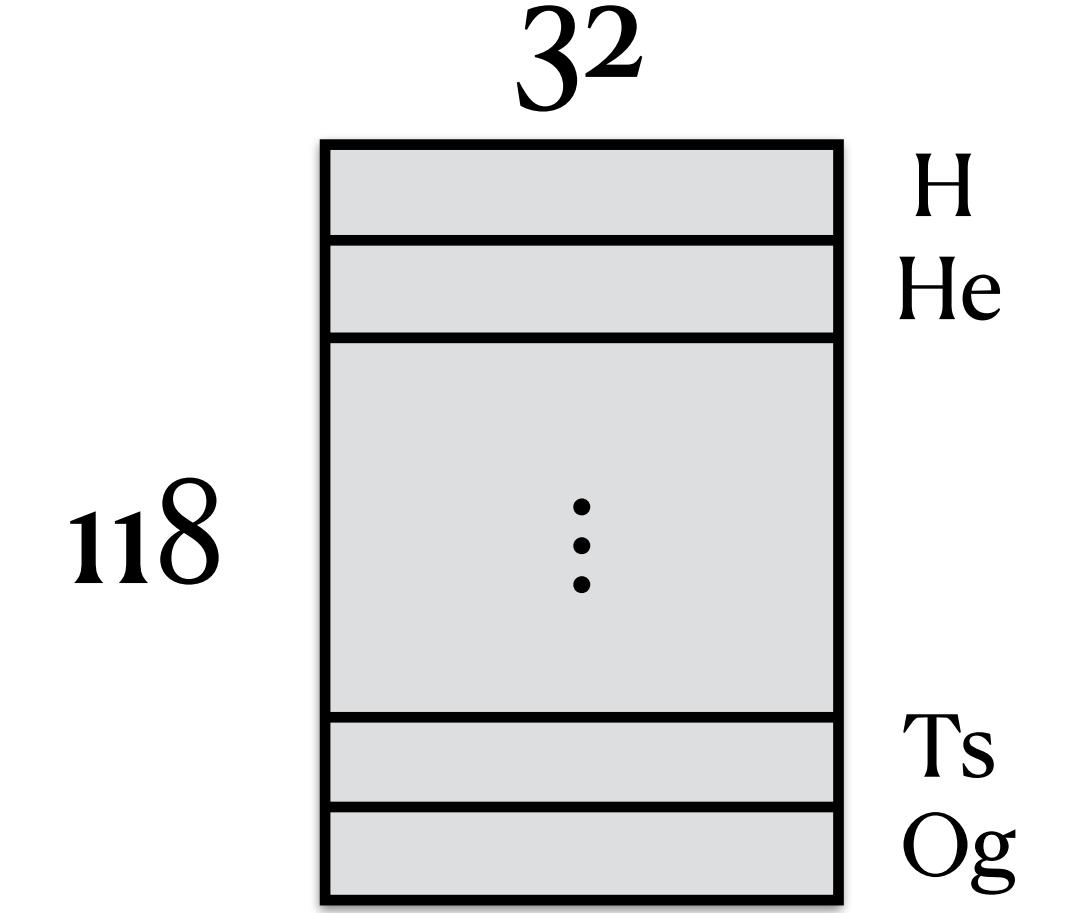
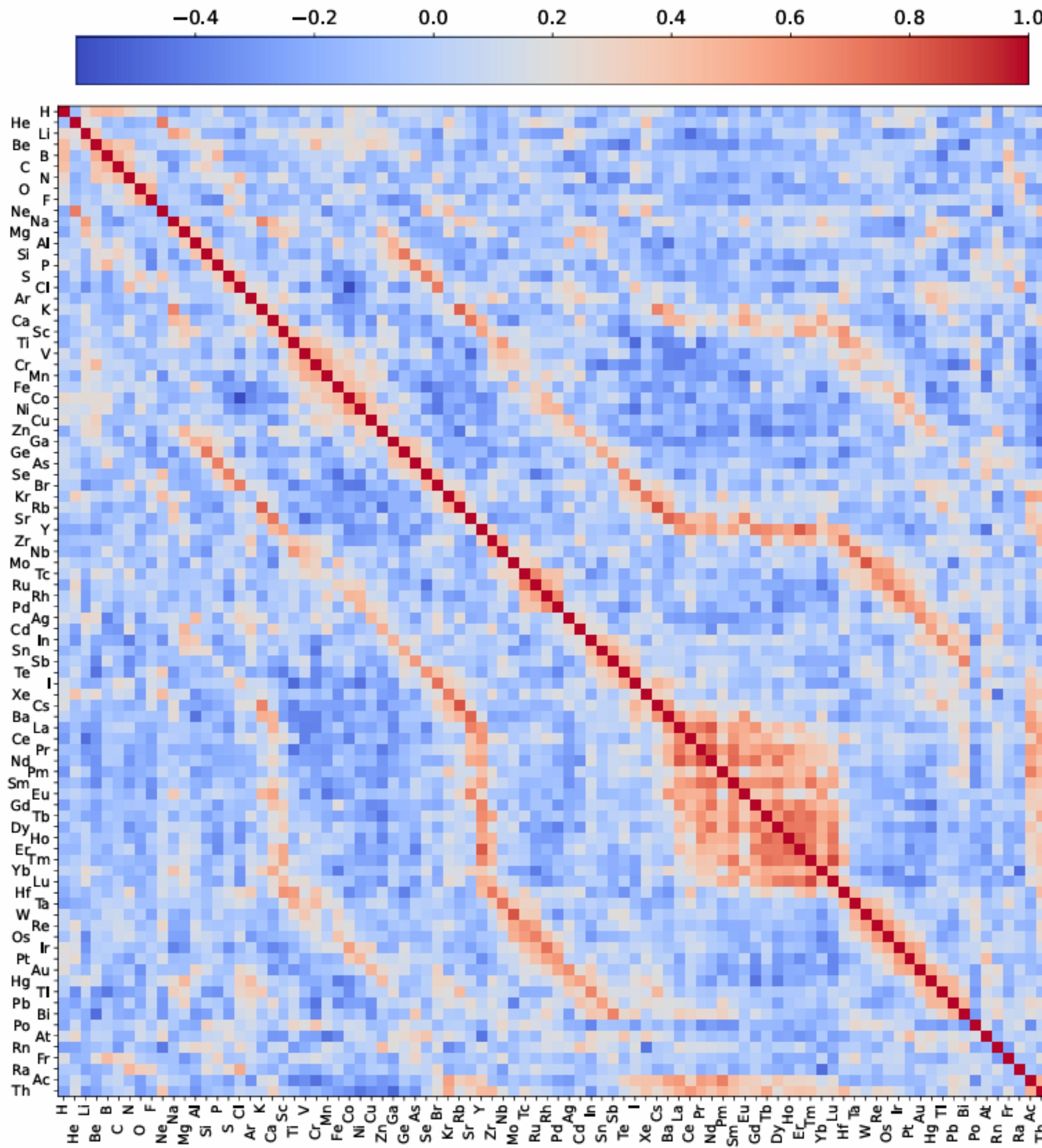
a-C-0-0-1/4-b-Ca-0-0-0-e-O-0.257-0-1/4-X-4.99-4.99-17.07-90-90-120

w_1 a_1 x_1 y_1 z_1 w_2 a_2 x_2 y_2 z_2 w_3 a_3 x_3 y_3 z_3 w_4 ℓ



167-a-C-0-0-1/4-b-Ca-0-0-0-e-O-0.257-0-1/4-X

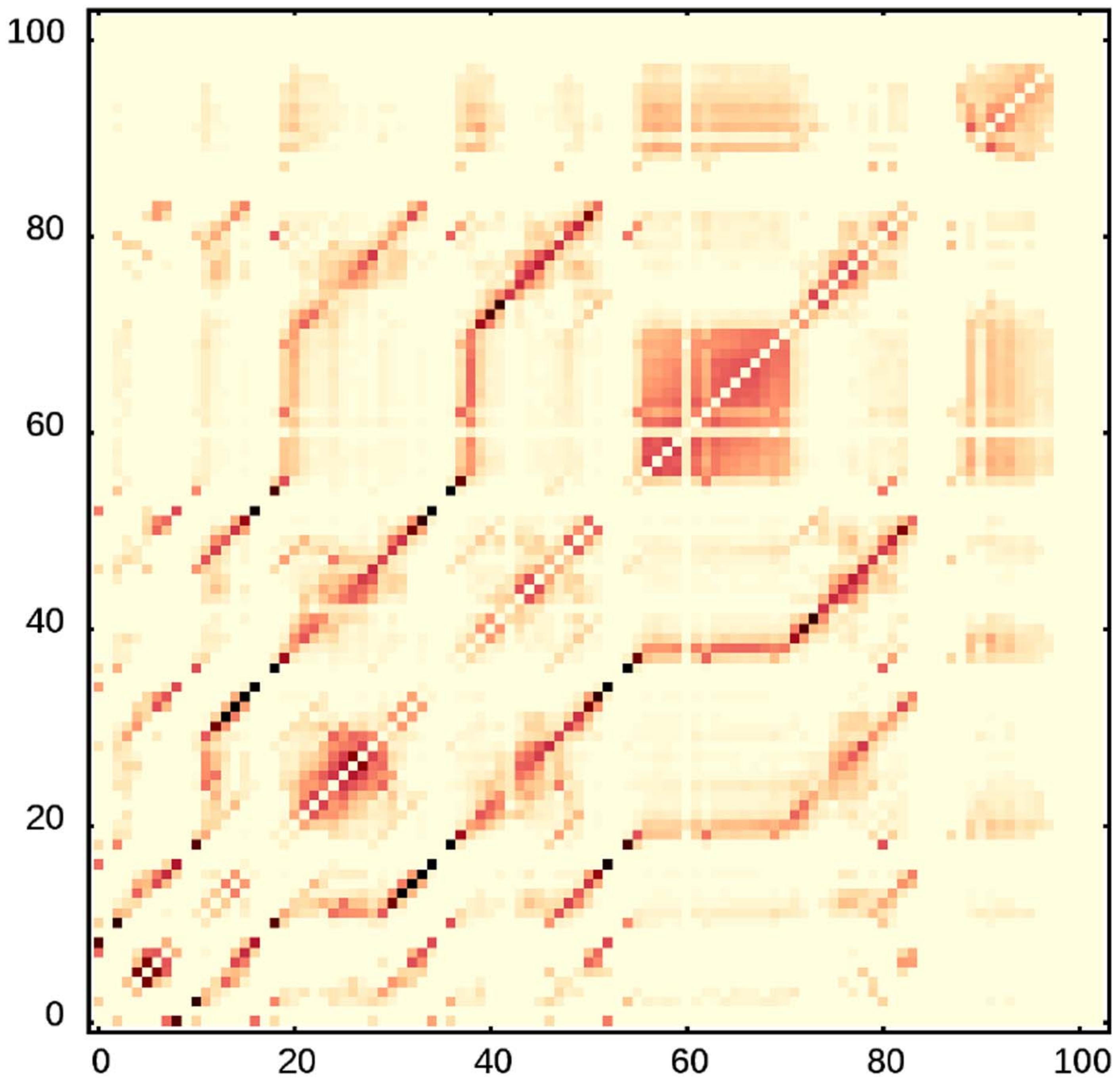




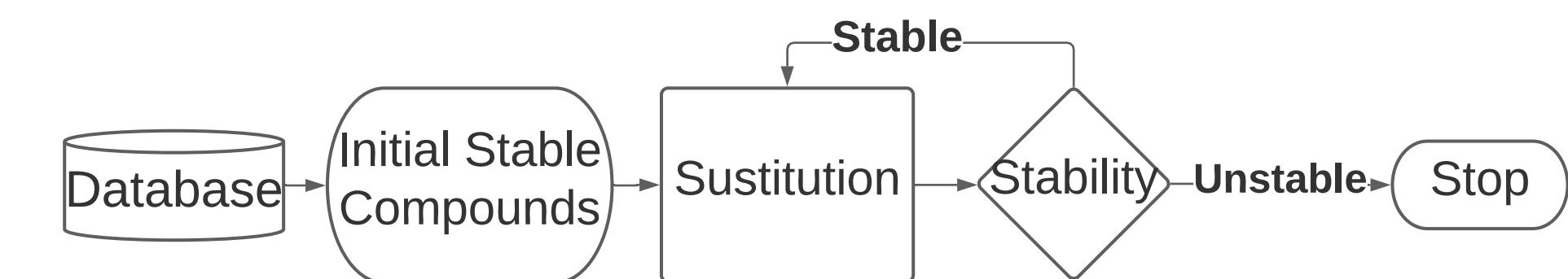
Element
embedding table

$$\frac{a \cdot b}{|a| \cdot |b|}$$

Cosine
similarity



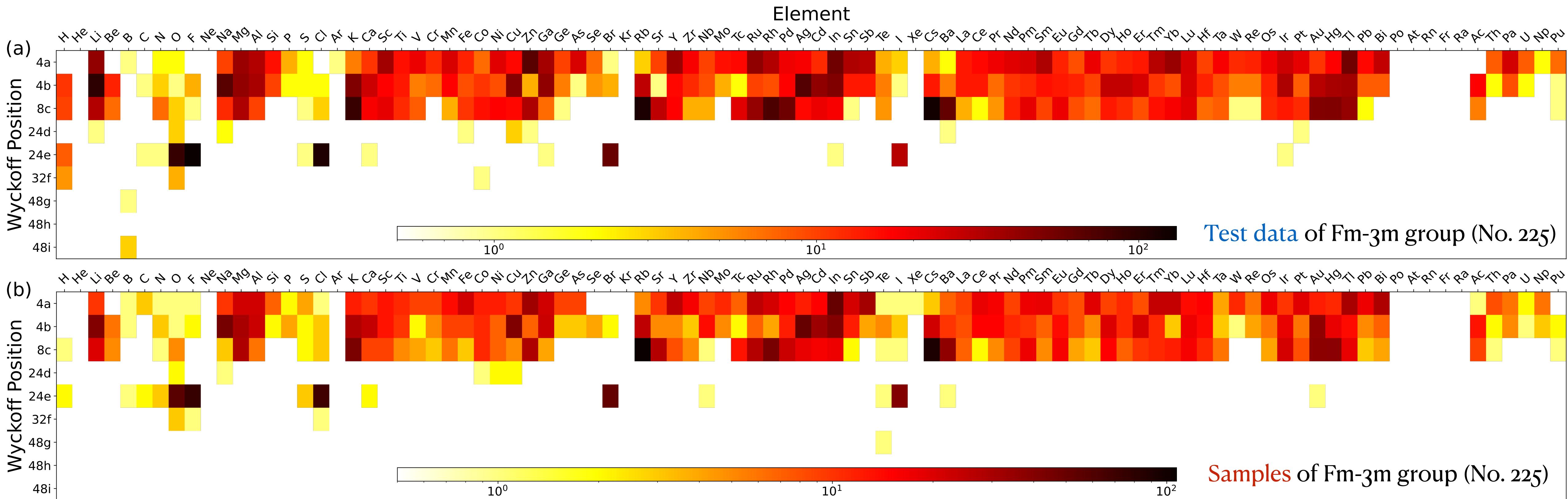
Purposely constructed
substitution partners based
on materials in ICSD



Glawe et al, NJP 2016
Wang et al, npj Comp. Mat. 2021

$W - A$ histogram

Where do elements go in solid state environment ?



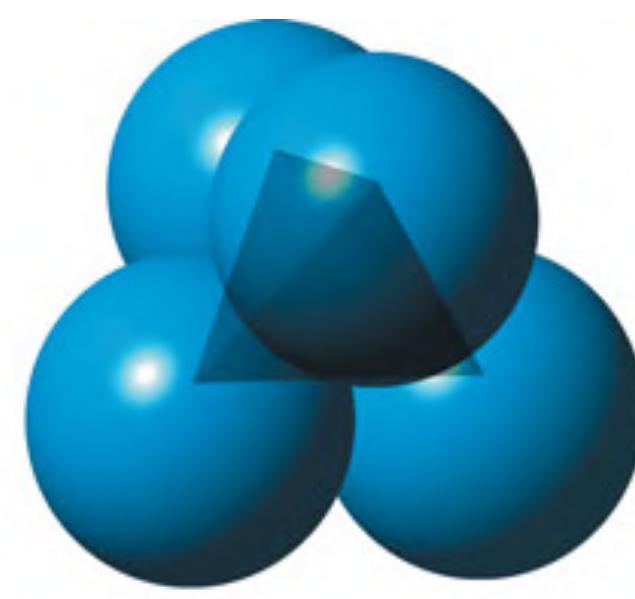
Nobel gas elements are inert in solids

Oxygen and halogen elements (F, Cl, Br, I) appear often at “24e” position

The model learns “idioms” of solid state chemistry for materials generation

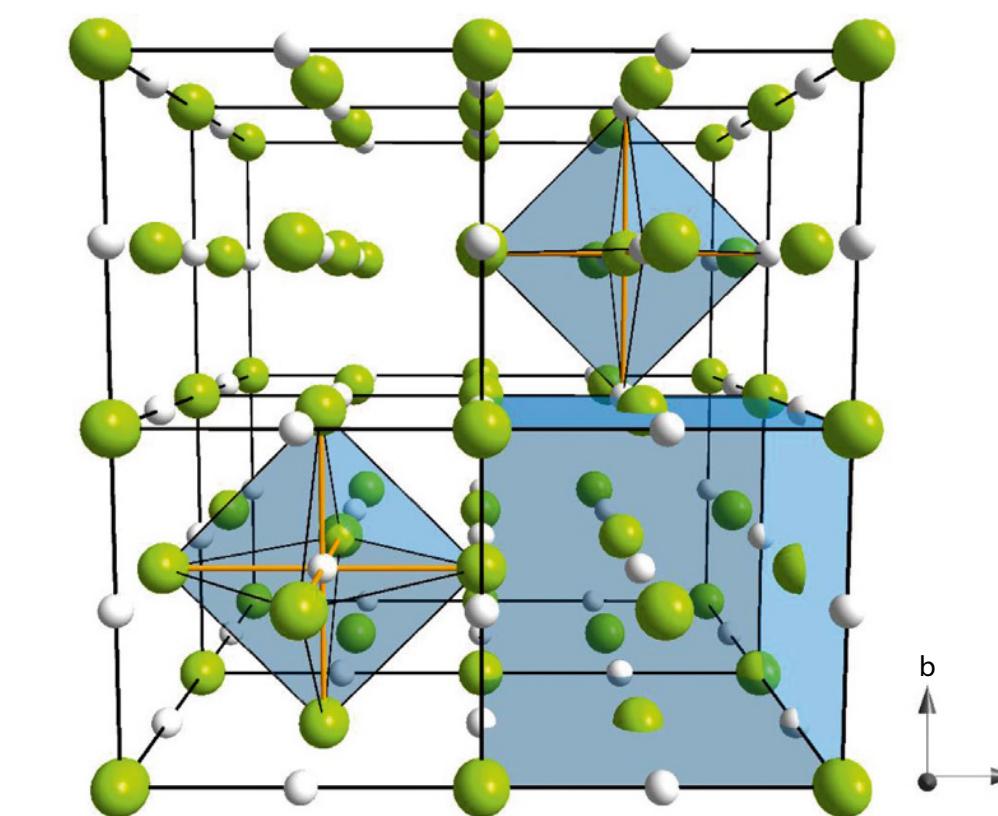
Validity test of crystals

Structure validity

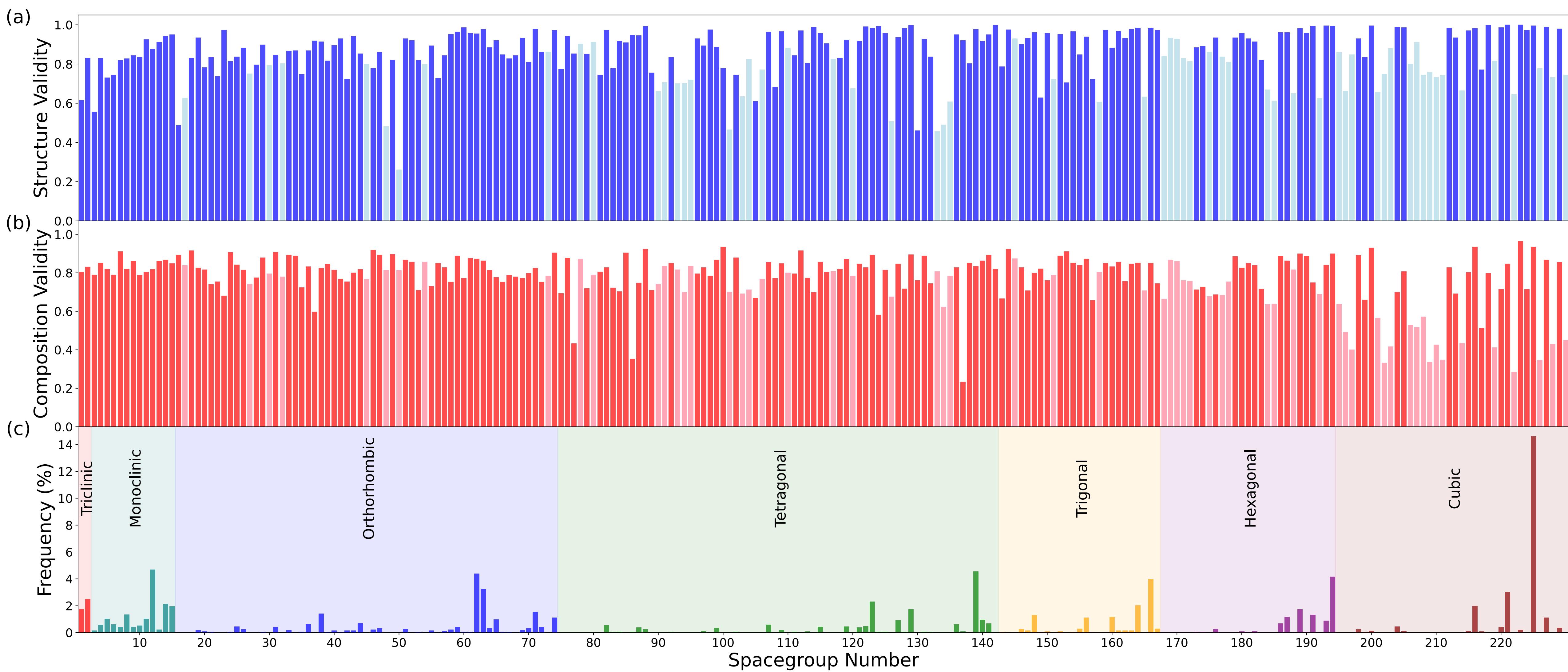


Atoms are not too close

Compositional validity



Charge balance



Space group	Crystal system	Training samples	Validity (%) ↑	
			Struc.	Comp.
2	Triclinic	676	83.10	83.0
12	Monoclinic	1273	87.70	81.80
62	Orthorhombic	1187	95.50	87.20
139	Tetragonal	1233	97.70	83.40
166	Trigonal	1076	98.50	85.0
194	Hexagonal	1129	99.40	89.90
225	Cubic	3960	99.60	93.50
1	Triclinic	27136	91.40	80.20

Autoregressive models

PGSchNet [38]	99.65	75.96
LM-CH (character-level tokenization) [17]	84.81	83.55
LM-AC (atom coordinate-level tokenization) [17]	95.81	88.87
Crystal-LLM [19] ¹	96.5	86.3

Diffusion models

CDVAE [11]	100.0	86.70
DiffCSP [12]	100.0	83.25
DiffCSP++ [24] ²	99.94	85.12
UniMat-Large [14]	97.2	89.4

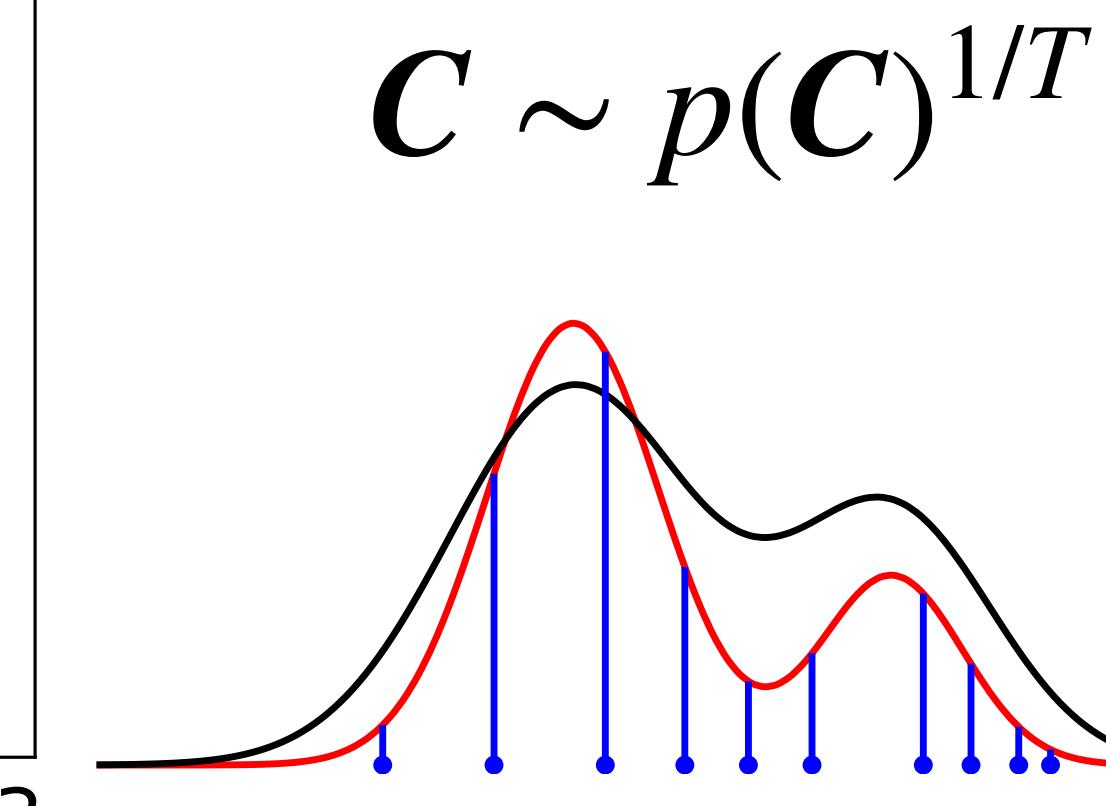
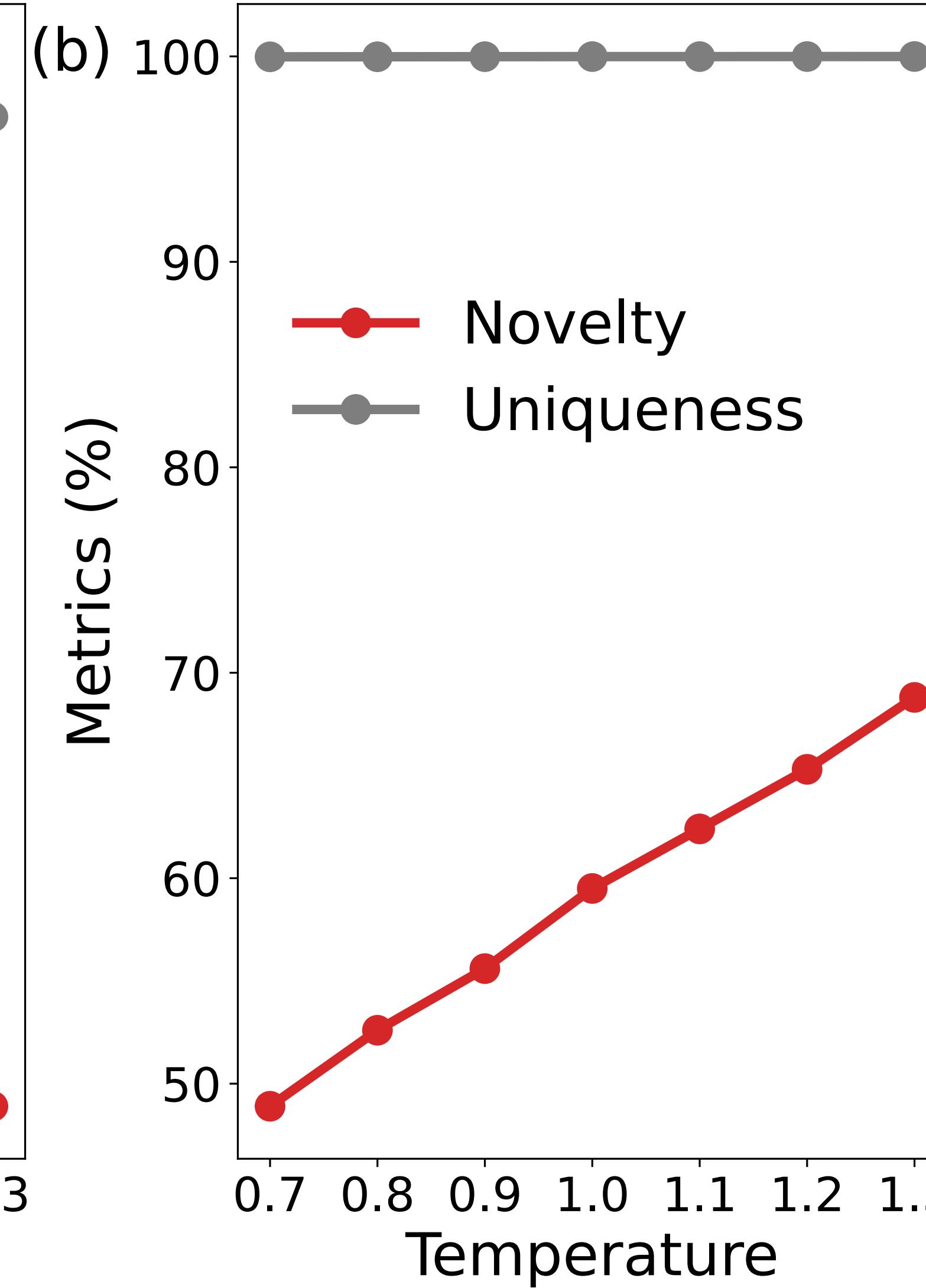
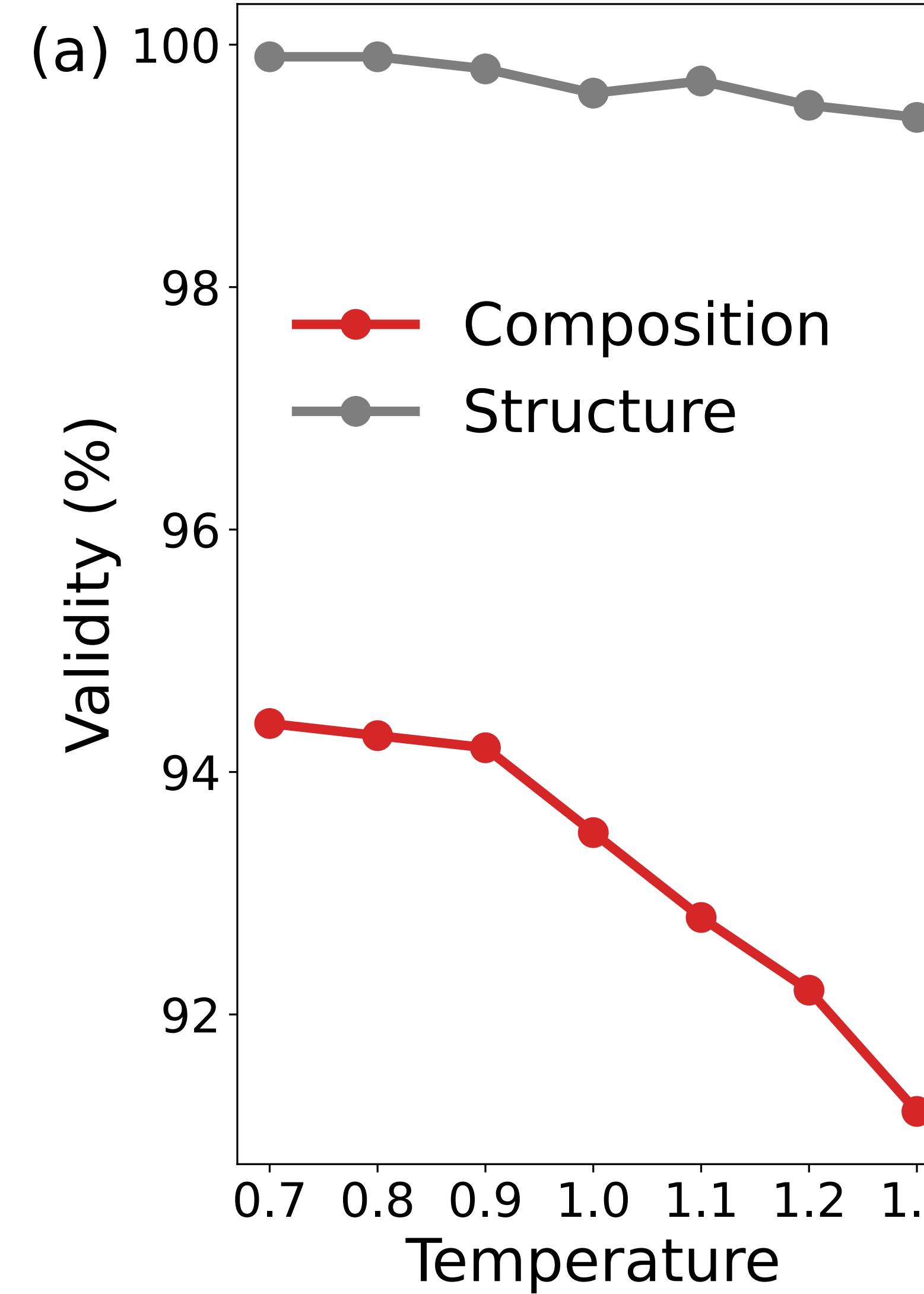
Higher validity scores for more symmetric space groups

Wyckoff position of those space groups are highly-structured, leaving few continuous DOF

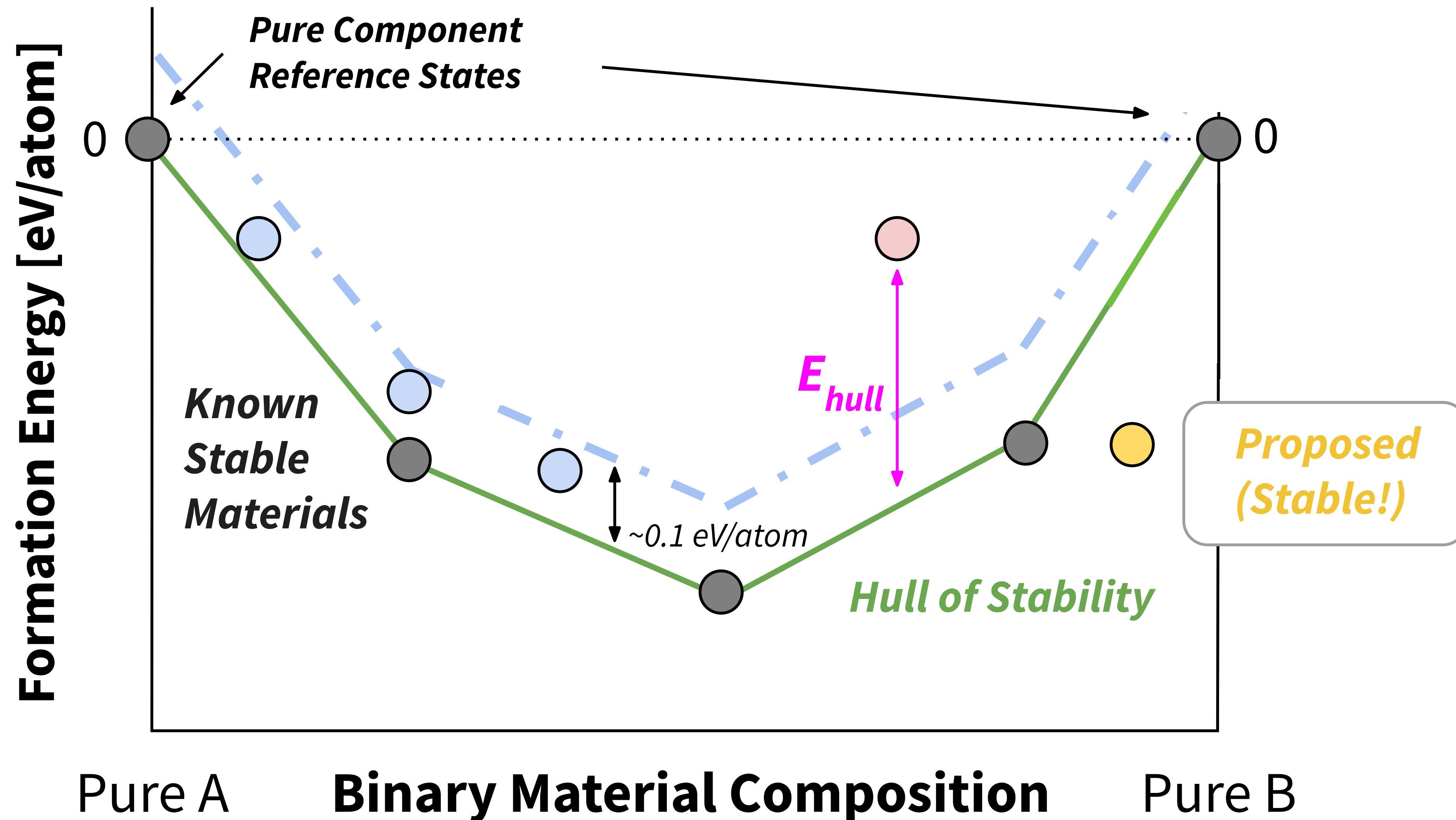
Generating those highly symmetric crystal samples are challenging for alternative approaches

Validity and diversity/novelty tradeoff

Fm-3m (No. 225) samples

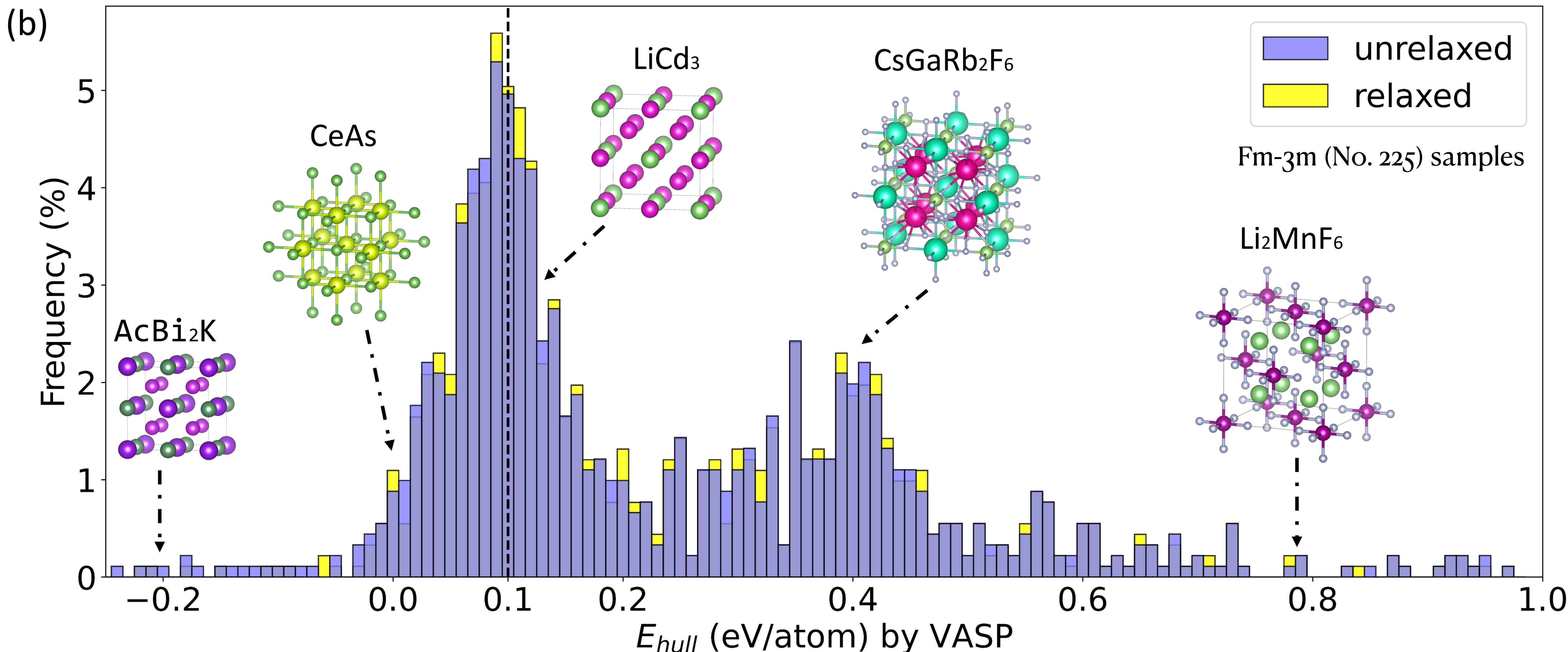


The stability test



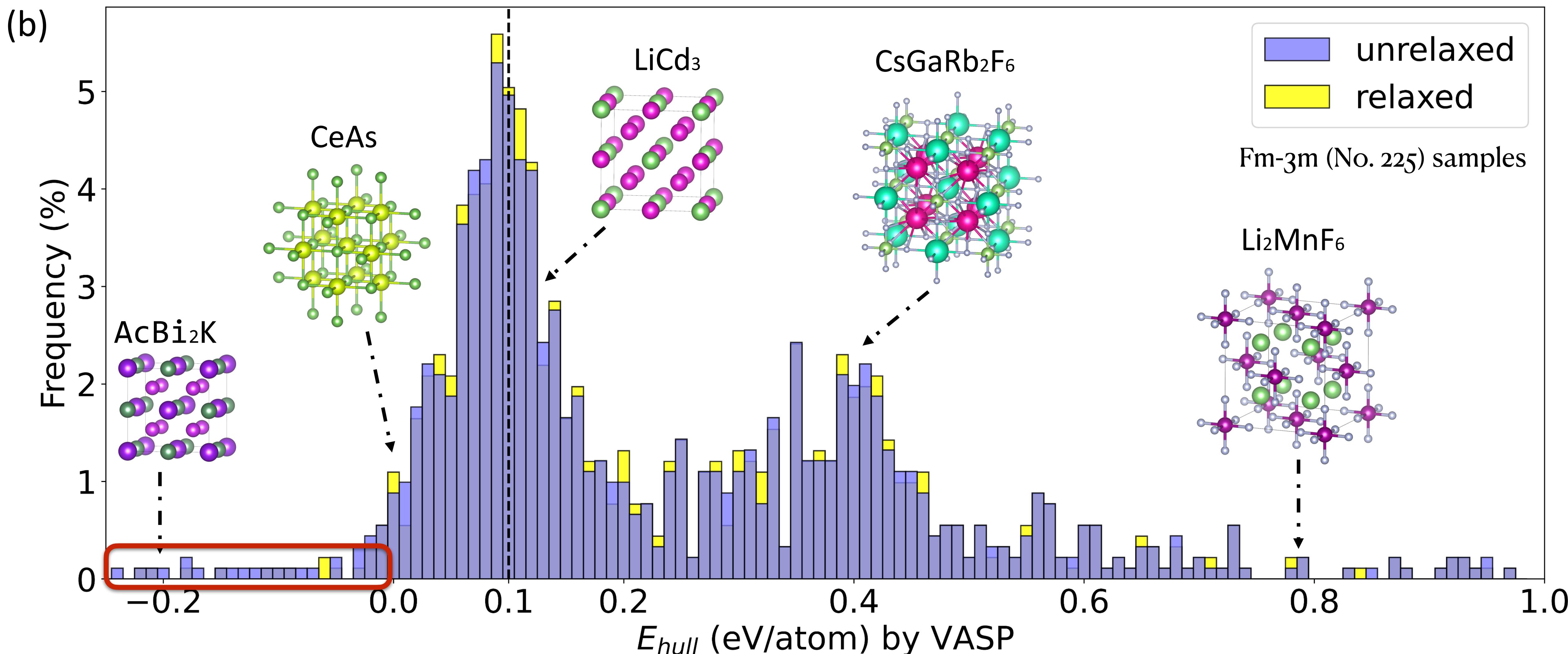
%30 of generated samples are within 0.1 eV/atom above hull

Relaxation do not change much, meaning that they were already very close to DFT local minima

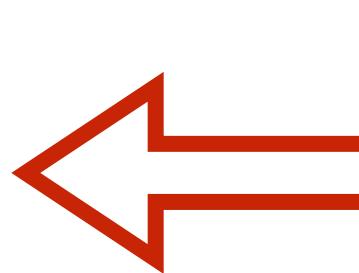


%30 of generated samples are within 0.1 eV/atom above hull

Relaxation do not change much, meaning that they were already very close to DFT local minima



In MP		Not in MP	
Formula	E_{hull} (eV/atom)	Formula	E_{hull} (eV/atom)
PdPtTm ₂	-0.00164	Ac ₂ CuRh	-0.14248
LiPt ₂ Zr	-0.01668	Ac ₂ AgSi	-0.04364
CdLi ₂ Pb	-0.00313	Ir ₂ LuPm	-0.0289
Li ₂ NdPb	-0.05703	InPm ₂ Tl	-0.01318
BeOs ₂ Si	-0.00779	PdPm ₂ Zn	-0.0843
CeAs	-0.0073	AuInPm ₂	-0.0568
Xe	-0.00067	AcBi ₂ K	-0.20215
InLiY ₂	-0.17504	InTb	-0.26278
InNd	-0.21393	Ac ₂ HgIn	-0.40448
Be ₂ C	-0.01227	CaPaRu ₂	-0.09344
EuTl	-0.1054		



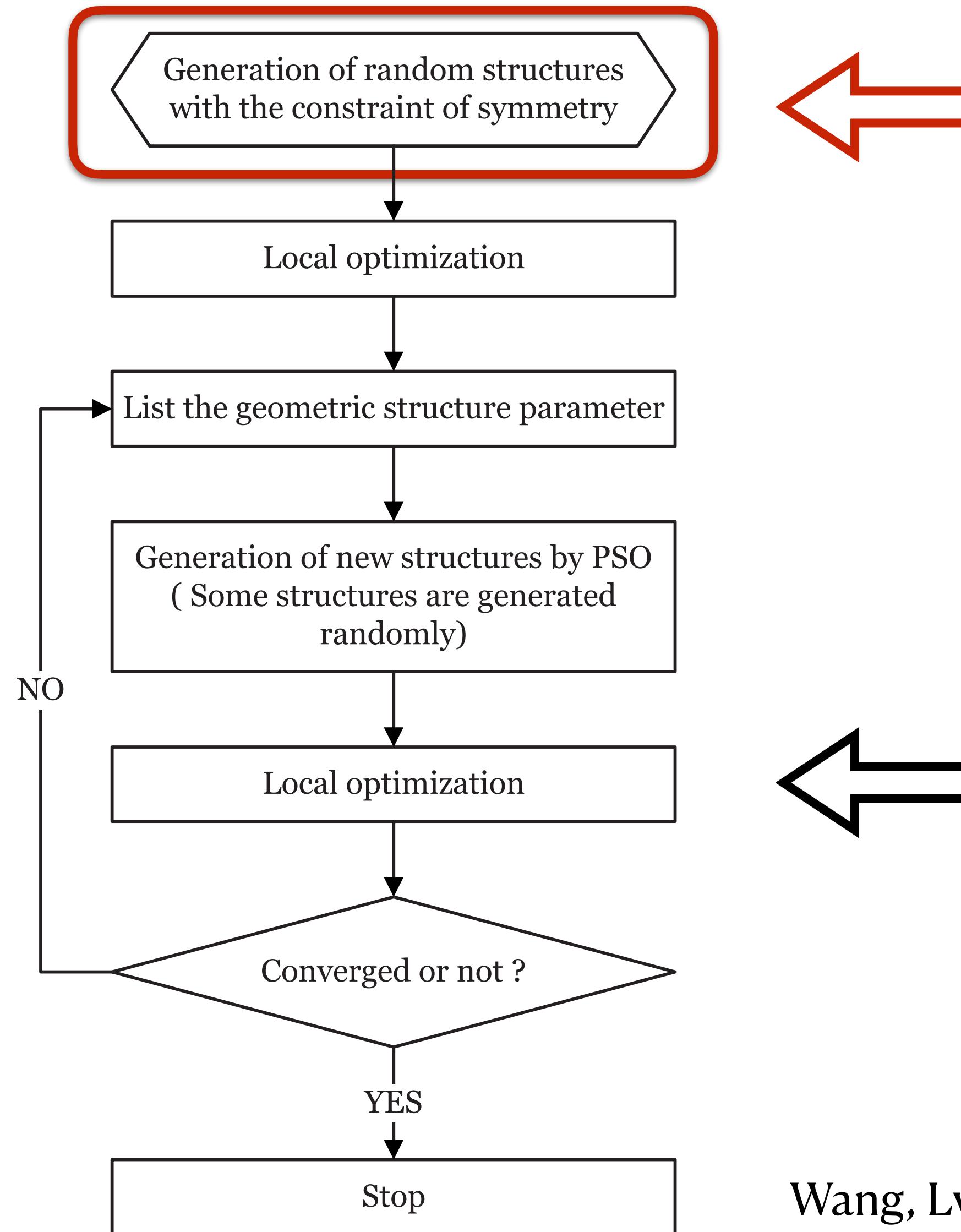
Discovered
stable materials

Cost of the discovery

- 13 GPU hours training the model
- 1 GPU minute sampling 1000 candidates
- 10 CPU hours DFT energy computation*

*can be accelerated by ML potential (DeepMD, MACE, M₃GNET, GNoME,...)

Applications: Structure initialization



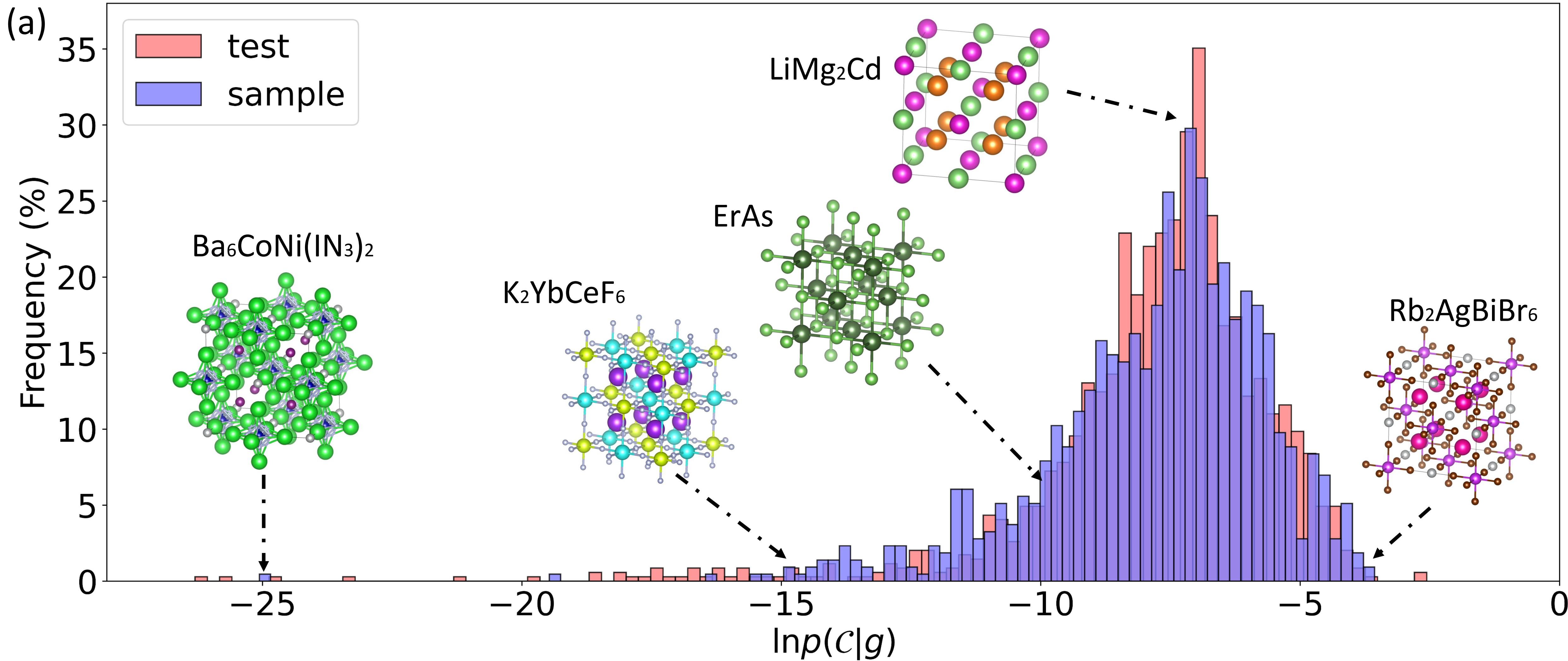
Drop-in replacement of
“random” symmetry structure generation
informative random, not uniform random

Surrogate machine learning models
focus on accelerating this step

Wang, Lv, Zhu, Ma, PRB 2010
USPEX, ARISS, ... are similar

FIG. 1. The flow chart of CALYPSO.

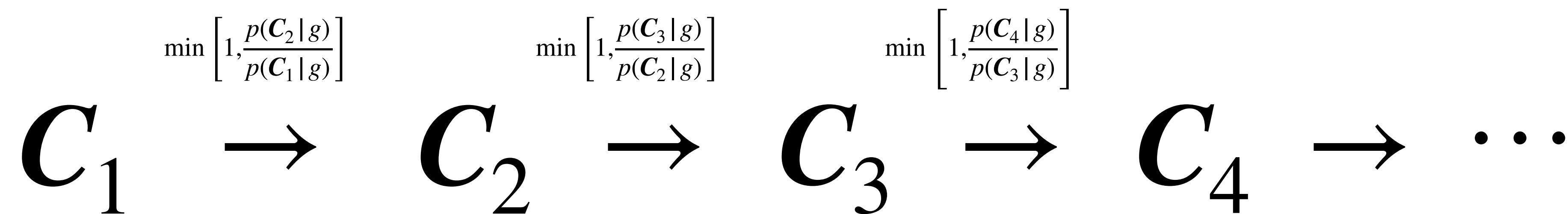
Likely, unlikely and typical Fm-3m crystals



Applications: Structure mutation

LW: What is your intuition for discovering new materials?

YGS (a colleague who grows crystals at IOP): Element substitution!



CrystalFormer guided Monte Carlo structure search

make it smarter

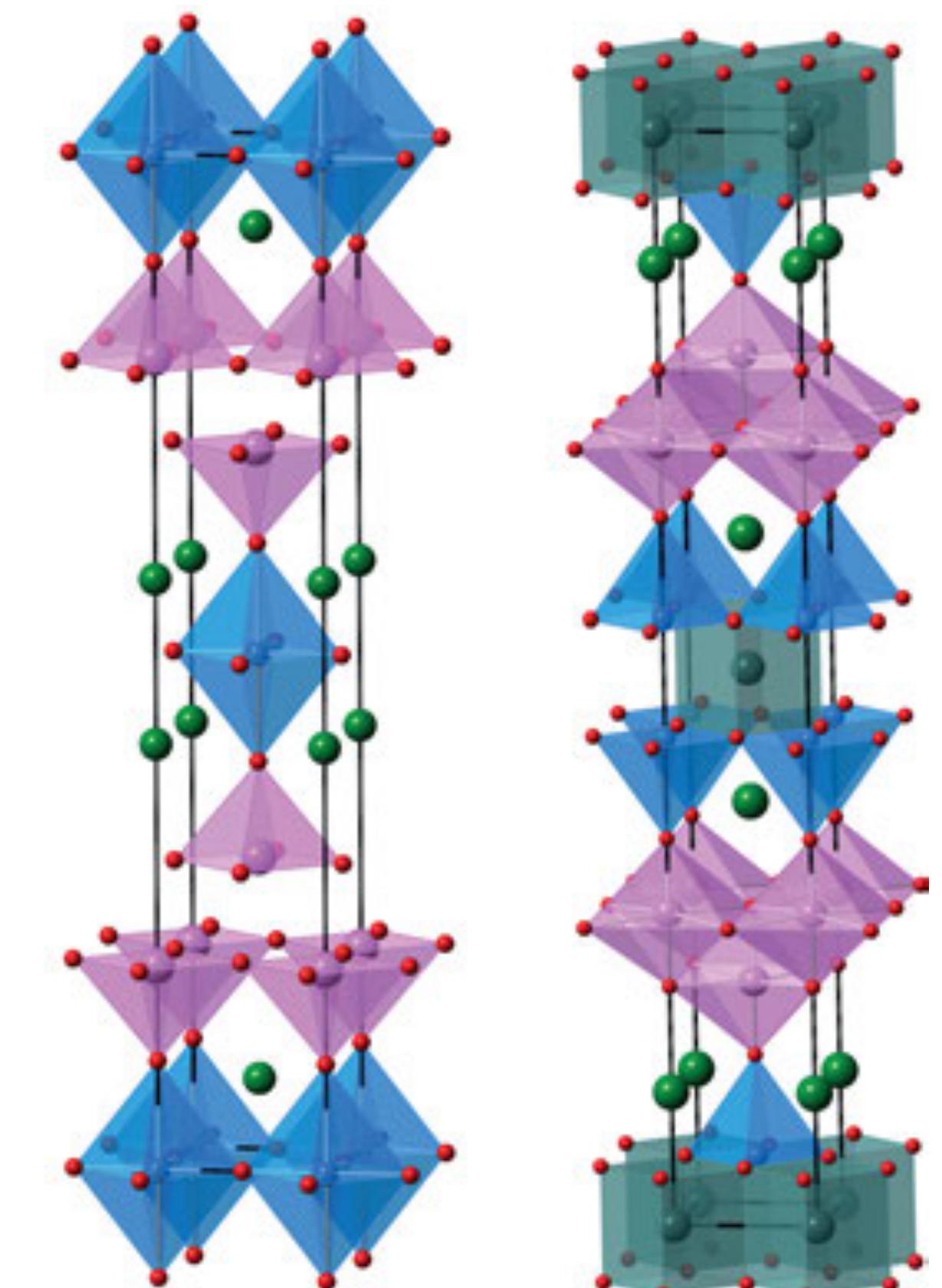
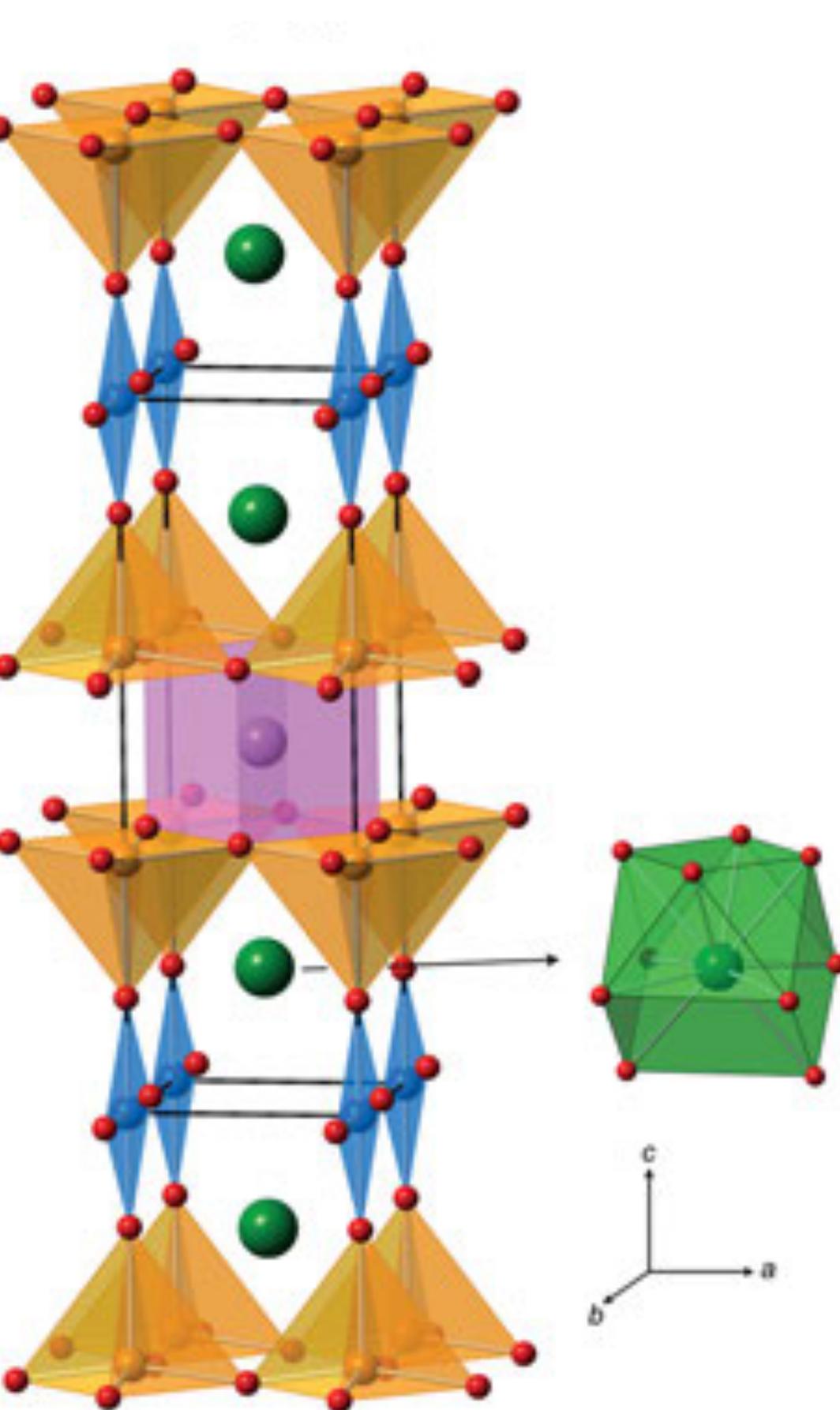
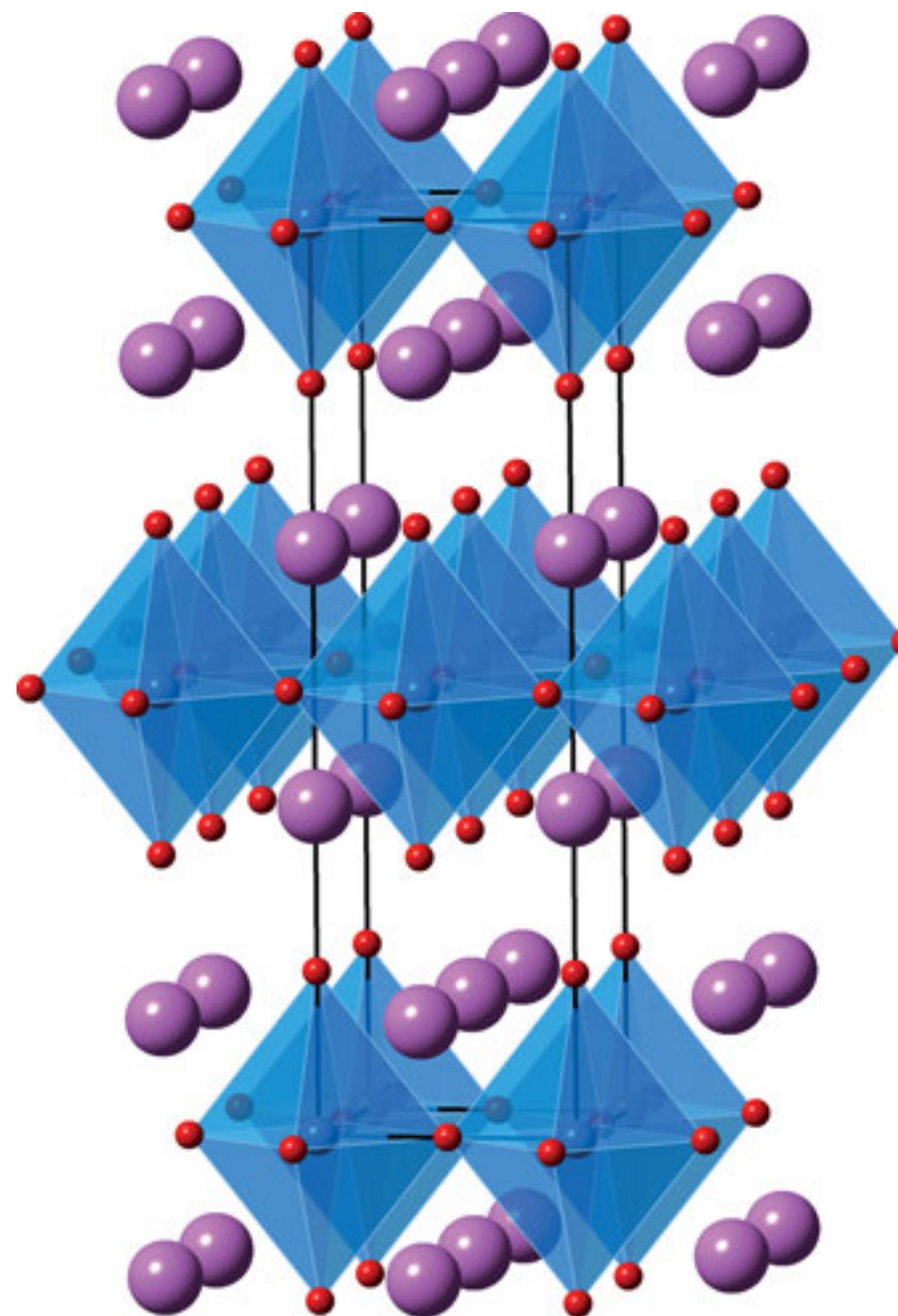


 **You**
make it even more smarter


 **ChatGPT**
We've reached a point where the depiction of the detective's intelligence is already at a transcendent, almost mythical level, merging with the universe's knowledge and represented as an ethereal entity of energy and information. Going beyond this might enter into realms of abstract concepts that are challenging to visualize concretely. However, if you have a specific idea or theme in mind that you believe would symbolize an even greater level of intelligence, I'd be happy to hear it and create an image based on your vision!

make T_c higher



Summary and Outlook

- *CrystalFormer* combines exact space group symmetry of crystallography and empirical chemistry from materials data
- Next: conditional generation based on desired properties (materials design) or experimental observations (solving structure)
- There is no obstruction towards a foundational model covers ALL inorganic crystalline materials in near future (< 1 year).